QUANTITATIVE INFRARED SPECTROSCOPY IN CHALLENGING ENVIRONMENTS: APPLICATIONS TO PASSIVE REMOTE SENSING AND PROCESS MONITORING

by

Qiaohan Guo

An Abstract

Of a thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Chemistry
in the Graduate College of
The University of Iowa

December 2012

Thesis Supervisor: Professor Gary W. Small

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **DEC 2012**	2. REPORT TYPE	3. DATES COVERED **00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE **Quantitative Infrared Spectroscopy in Challenging Environments: Applications to Passive Remote Sensing and Process Monitoring**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Iowa,Iowa City,IA,52242**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S)
	11. SPONSOR/MONITOR’S REPORT NUMBER(S)
12. DISTRIBUTION/AVAILABILITY STATEMENT **Approved for public release; distribution unlimited**	
13. SUPPLEMENTARY NOTES	

14. ABSTRACT

Chemometrics is a discipline of chemistry which uses mathematical and statistical tools to help in the extraction of chemical information from measured data. With the assistance of chemometric methods, infrared (IR) spectroscopy has become a widely applied quantitative analysis tool. This dissertation explores two challenging applications of IR spectroscopy facilitated by chemometric methods: (1) passive Fourier transform (FT) remote sensing and (2) process monitoring by near-infrared (NIR) spectroscopy. Passive FT-IR remote sensing offers a measurement method to detect gaseous species in the outdoor environment. Two major obstacles limit the application of this method in quantitative analysis: (1) the effect of both temperature and concentration on the measured spectral intensities and (2) the difficulty and cost of collecting reference data for use in calibration. To address these problems, a quantitative analysis protocol was designed based on the use of a radiance model to develop synthetic calibration data. The synthetic data served as the input to partial least-squares (PLS) regression in order to construct models for use in estimating ethanol and methanol concentrations. The methodology was tested with both laboratory and field remote sensing data. Near-infrared spectroscopy has attracted significant interest in process monitoring because of the simplicity in sample preparation and the compatibility with aqueous solutions. For use in process monitoring, the need exists for robust calibrations. A challenge in the NIR region is that weak, broad and highly overlapped spectral bands make it difficult to extract useful chemical information from measured spectra. In this case, signal processing methods can be helpful in removing unwanted signals and thereby uncovering useful information. When applying signal processing as a spectral preprocessing tool and regression analysis for building a quantitative calibration model, optimizing the parameters that specify the details of the methods is crucial. In this research, particle swarm optimization, a population-based optimization method was applied. Digital filtering and wavelet processing methods were evaluated for their utility as spectral preprocessing tools. Both a pump-controlled flowing system and bioreactor runs involving the yeast Pichia pastoris, were studied in this work. In investigating the bioreactor runs, insufficient reference data resulted in difficulties in employing the PLS calibration method. Instead, the augmented classical least-squares modeling technique was applied since it requires only pure-component or composite spectra of the analyte and background matrix rather than a large set of mixture samples of known analyte concentration.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**259**	

ABSTRACT

Chemometrics is a discipline of chemistry which uses mathematical and statistical tools to help in the extraction of chemical information from measured data. With the assistance of chemometric methods, infrared (IR) spectroscopy has become a widely applied quantitative analysis tool. This dissertation explores two challenging applications of IR spectroscopy facilitated by chemometric methods: (1) passive Fourier transform (FT) remote sensing and (2) process monitoring by near-infrared (NIR) spectroscopy.

Passive FT-IR remote sensing offers a measurement method to detect gaseous species in the outdoor environment. Two major obstacles limit the application of this method in quantitative analysis: (1) the effect of both temperature and concentration on the measured spectral intensities and (2) the difficulty and cost of collecting reference data for use in calibration. To address these problems, a quantitative analysis protocol was designed based on the use of a radiance model to develop synthetic calibration data. The synthetic data served as the input to partial least-squares (PLS) regression in order to construct models for use in estimating ethanol and methanol concentrations. The methodology was tested with both laboratory and field remote sensing data.

Near-infrared spectroscopy has attracted significant interest in process monitoring because of the simplicity in sample preparation and the compatibility with aqueous solutions. For use in process monitoring, the need exists for robust calibrations. A challenge in the NIR region is that weak, broad and highly overlapped spectral bands make it difficult to extract useful chemical information from measured spectra. In this case, signal processing methods can be helpful in removing unwanted signals and thereby uncovering useful information. When applying signal processing as a spectral preprocessing tool and regression analysis for building a quantitative calibration model, optimizing the parameters that specify the details of the methods is crucial. In this research, particle swarm optimization, a population-based optimization method was applied. Digital filtering and wavelet processing methods were evaluated for their utility as spectral preprocessing tools.

Both a pump-controlled flowing system and bioreactor runs involving the yeast, *Pichia pastoris*, were studied in this work. In investigating the bioreactor runs, insufficient reference data resulted in difficulties in employing the PLS calibration method. Instead, the augmented classical least-squares modeling technique was applied since it requires only pure-component or composite spectra of the analyte and background matrix rather than a large set of mixture samples of known analyte concentration.

Abstract Approved:

Thesis Supervisor

Title and Department

Date

QUANTITATIVE INFRARED SPECTROSCOPY IN CHALLENGING ENVIRONMENTS: APPLICATIONS TO PASSIVE REMOTE SENSING AND PROCESS MONITORING

by

Qiaohan Guo

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Chemistry
in the Graduate College of
The University of Iowa

December 2012

Thesis Supervisor: Professor Gary W. Small

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Qiaohan Guo

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Chemistry at the December 2012 graduation.

Thesis Committee: ______________________________
Gary W. Small, Thesis Supervisor

____
Mark A. Arnold

____
Lei Geng

____
James B. Gloer

____
Julie L.P. Jessop

ACKNOWLEDGEMENTS

This work would not have been possible without the support of many people. I gratefully thank my advisor, Prof. Gary W. Small, for guiding me on observing the world of science, directing me on research, and always encouraging me to spread my wings. I am also grateful to my committee members, Prof. Mark A. Arnold, Prof. Lei Geng, Prof. James B. Gloer and Prof. Julie J.P. Jessop. They offered invaluable discussion and suggestions on my work. Many thanks for the important experiences and kind help from former group members, Dr. Yusuf I. Sulub, Dr. Boyong Wan, Dr. Yuping Wu, Zongqin Ruan, Yingqing Wang, Wei Wang as well as current group members, Chamathca Priyanwada, Sanjeewa Rasika Karunathilaka, Hua Yu and Brian Dess.

The data collection and funding in the remote sensing work was supported by the Department of Energy, U.S. Environmental Protection Agency, Department of the Army and AeroSurvey, Inc. Dr. Robert T. Kroutil and Dr. Roger Combs are acknowledged for the laboratory data collection in this project. I thank Dr. Mark A. Arnold and his research group for data collection in the near infrared study, as well as Dr. Elizabeth Gibson and Dr. Jonathon Olesberg for data collection.

At last, I would like to express my deepest gratitude to my parents. Nothing could have been achieved without their love, support and encouragement throughout my life and especially the graduate study carried out in the US.

ABSTRACT

Chemometrics is a discipline of chemistry which uses mathematical and statistical tools to help in the extraction of chemical information from measured data. With the assistance of chemometric methods, infrared (IR) spectroscopy has become a widely applied quantitative analysis tool. This dissertation explores two challenging applications of IR spectroscopy facilitated by chemometric methods: (1) passive Fourier transform (FT) remote sensing and (2) process monitoring by near-infrared (NIR) spectroscopy.

Passive FT-IR remote sensing offers a measurement method to detect gaseous species in the outdoor environment. Two major obstacles limit the application of this method in quantitative analysis: (1) the effect of both temperature and concentration on the measured spectral intensities and (2) the difficulty and cost of collecting reference data for use in calibration. To address these problems, a quantitative analysis protocol was designed based on the use of a radiance model to develop synthetic calibration data. The synthetic data served as the input to partial least-squares (PLS) regression in order to construct models for use in estimating ethanol and methanol concentrations. The methodology was tested with both laboratory and field remote sensing data.

Near-infrared spectroscopy has attracted significant interest in process monitoring because of the simplicity in sample preparation and the compatibility with aqueous solutions. For use in process monitoring, the need exists for robust calibrations. A challenge in the NIR region is that weak, broad and highly overlapped spectral bands make it difficult to extract useful chemical information from measured spectra. In this case, signal processing methods can be helpful in removing unwanted signals and thereby uncovering useful information. When applying signal processing as a spectral preprocessing tool and regression analysis for building a quantitative calibration model, optimizing the parameters that specify the details of the methods is crucial. In this research, particle swarm optimization, a population-based optimization method was applied. Digital filtering and wavelet processing methods were evaluated for their utility as spectral preprocessing tools.

Both a pump-controlled flowing system and bioreactor runs involving the yeast, *Pichia pastoris*, were studied in this work. In investigating the bioreactor runs, insufficient reference data resulted in difficulties in employing the PLS calibration method. Instead, the augmented classical least-squares modeling technique was applied since it requires only pure-component or composite spectra of the analyte and background matrix rather than a large set of mixture samples of known analyte concentration.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1
INTRODUCTION

Analytical chemistry is an important branch of chemistry. The accurate and precise measurement of chemical species is necessary in a variety of settings. Qualitative and quantitative examinations of chemical samples have been widely incorporated into various fields of chemistry research, as well as clinical, environmental, pharmaceutical, industrial and forensic applications. With improvements in instrumentation and electronics, the available chemical measurement methodologies are increasingly precise and accurate. With the advent of powerful laboratory computers and computer-controlled instrumentation, the analyst has the ability to acquire data in ever increasing amounts. Data can be acquired faster and in quantities that exceed the capacity of the analyst to perform visual or manual data interpretation. As a consequence, an important focus of current research efforts is to develop automated or semi-automated techniques to aid in converting raw experimental data to useful chemical information.

1.1 Chemometrics

Chemometrics is a subdiscipline of analytical chemistry. In 1974, the Chemomtrics Society was founded in Seattle, Washington. Chemometrics is defined as the development and application of mathematical and statistical methods to help chemists perform better experimental designs and improve their ability to interpret and extract useful information (e.g., knowledge of the constitution or reaction mechanism) from chemical measurements. A further emphasis is to enable these goals to be accomplished rapidly or in an automated manner.

Chemometrics is not the only science which facilitates good experimental procedures and information extraction. Psychometrics, biometrics and econometrics are the formal areas of study in psychology, biology and economics that focus on related goals.[1,2] The development of computerized analytical instrumentation has enabled the rapid growth of chemometrics. With increasing demands for qualitative and quantitative analysis tools, chemometrics has been widely incorporated into all branches of chemistry, especially analytical chemistry. For example, chromatography, spectroscopy and mass spectrometry methods have all derived benefits from developments in chemometrics.[3] Research in this thesis will focus on the development of chemometric methods for use with applications of IR spectroscopy.

1.2 Infrared Spectroscopy

Infrared spectroscopy studies the interaction between chemical species and the electromagnetic spectrum in the IR region. The energy of the radiation in the IR region can cause molecular vibrational and rotational motions. The motions and their associated energy levels are determined by the configuration and number of atoms within a molecule. A nonlinear shaped molecule containing N atoms possesses $3N-6$ fundamental vibrational modes, whereas a linear molecule can vibrate in $3N-5$ modes. Transitions from one level to another occur when there is a change in the dipole moment during the vibration. Based on this requirement, homonuclear diatomic molecules (e.g., N_2, O_2, H_2) are IR inactive.

As a result of the fundamental vibration, a functional group of a molecule corresponds to a particular radiated energy. The vibrations at different frequencies can reflect the characteristics and structure of a molecule, thereby creating a signature of a molecule if the individual frequencies across the IR region are interrogated (i.e., a spectrum is recorded).

In addition to the fundamental vibrational modes, the observed spectra also exhibit overtones and combinations of the corresponding bands. Hence, IR spectrometry can be a useful tool in structure investigation and the identification of an unknown sample. Furthermore, the intensity of the spectral signature is related to the abundance of the IR-active species in a sample. Therefore, both qualitative and quantitative analysis can be performed based on an IR spectroscopic measurement.[4]

1.3 Instrumentation

The most commonly used instrumentation in IR spectroscopy is the Fourier transform (FT) type. The heart of an FT-IR spectrometer is the interferometer developed by Michelson in the 1880's. With the development of computational capabilities that made possible the routine calculation of the FT, this type of instrument became popular during the 1970's. The first commercial FT-IR spectrometer was produced in this decade.

Compared to the traditional dispersive instrument, there are several advantages that an FT-IR instrument possesses. Instead of a wavelength scan, each interferometer scan generates an interferogram, which contains the information of the entire spectrum acquired simultaneously. The scan is fast and offers an excellent signal-to-noise ratio (SNR). Instead of prisms or diffraction gratings and their associated entrance and exit slits that limit light throughput, the FT-IR instrument uses beamsplitters and mirrors in directing the full-size light beam throughout the optical path. This can help maintain a high light throughput. Meanwhile, the scanning method is capable of high spectral resolution without significant loss of optical throughput and high wavelength precision.[5]

However, the primary drawback of the FT-IR instrument is the lack of ruggedness associated with the need for moving parts within the interferometer. This may present challenges in using the instrumentation in harsh or demanding environments such as a manufacturing plant. A more rugged spectrometer based on an acousto-optic tunable filter (AOTF) was employed in part of the data collection in this research. The AOTF is a type of electronic optical filter. The light scanning is performed based on the acousto-optic effect of the specially cut crystal. In the AOTF instrument, changing a radio frequency input to the crystal can cause a change in refractive index, thereby allowing light diffraction at different wavelengths.[6] A detailed introduction of IR spectroscopy, the associated FT-IR and AOTF instrumentation, and the related mathematics will be provided in Chapter 2. The specific experimental setups will be described in Chapters 4 and 5.

1.4 Applications

1.4.1 Environmental Monitoring

Air pollution is a serious environmental problem which can cause harm or discomfort to humans or other living organisms, or cause damage to the natural environment. Because of human activity (e.g., industrial releases, green house effect) or natural phenomena (e.g., volcanoes), the primary pollutants include sulfur oxides (SO_x), nitrogen oxides (NO_x), carbon dioxide (CO_2), carbon monoxide (CO) and other volatile organic compounds. A great effort has been made to develop fast and reliable detection of constituents of the atmosphere. Because of the high SNR and fast-scanning properties of FT instrumentation and the specificity of the mid-IR region, FT-IR spectrometry has gained popularity in remote gas detection.

Depending on the IR source, the technique of FT-IR remote sensing is divided into active and passive modes. The active mode has a stable, well controlled instrumental light source in the optical path. This approach has been applied to monitor the pollutant emission rates from coal mines[7,8], and generate spectral data on hazardous air pollutants[9–12]. Although the light source is stable in the active mode, the greatest difficulty in remote detection is to couple the light source into the spectrometer. To make the measurement, the gaseous sample has to be located in the optical path, i.e., between the light source and the spectrometer. The need for a controlled light source restricts the flexibility and mobility of the experimental setup. Consequently, the active mode is only applicable to detection in a fixed location, and not amenable to a moving scan, e.g., detection from a moving platform such as an aircraft.

The passive measurement mode has a more practical and simpler instrumental setup. Instead of a stable IR source, the passive detection views an uncontrolled IR radiant background. While this setup helps to simplify the instrument, it raises limitations and problems at the same time. First, based on the governing radiance model of the experiment, it requires a significant temperature difference between the background and the gas sample. This requirement results in a low sensitivity when the temperature difference is low. In addition, the naturally occurring background lacks stability, caused by either temperature changes or changes in the background scene within the instrumental field-of-view. This translates to difficulties in implementing the conventional calculation to obtain the analyte absorbance or emission by taking the ratio of the sample spectrum to a previously acquired background spectrum. Moreover, the unstable background brings in challenges regarding how to perform a reliable calibration of the measurement. Third, data collection in remote sensing measurements is expensive because of the equipment required and logistics associated with measurements made in the outdoor environment. Especially for toxic gas detection, collecting a large amount of calibration data outdoors is not practical. As a result, calibration is difficult in passive remote sensing. Strategies to overcome these calibration difficulties have been investigated. Classification methods by calibration transfer from either ground or synthetic data to field data collected from the air have been studied in our laboratory.[13,14]

In this thesis, a quantitative method is developed based on the mathematical synthesis of the calibration data by appropriate radiance models. The radiance models will be introduced in Chapter 2. Chapter 4 will introduce the detailed protocol and procedure for the synthetic calibration method.

1.4.2 Process Monitoring

Besides passive remote sensing, process monitoring is also an important field of application for IR spectroscopy. As a replacement for periodic discrete sampling, continuous monitoring is more desirable due to increasing demands in industry for real-time response and quality control. Near-infrared (NIR) spectroscopy has been investigated in continuous monitoring applications because of the nondestructive nature of the measurement and the ease of sample preparation. In addition, the lower water absorbance in the NIR region relative to the mid-IR makes it compatible with aqueous sampling environments. NIR measurements have been in applied in various aspects of industrial and pharmaceutical study, including the monitoring of film coatings,[15] blending and its corresponding kinetics,[16,17] fermentation processes,[18] and polymer extrusion.[19,20]

The primary challenge for NIR measurement applications lies in the area of data analysis and interpretation because the spectral features in this region are relatively broad and weak. The presence of overlapping spectral features makes it difficult to extract the signal associated with a target analyte from the overlapping spectral background. Therefore, chemometric techniques are important to enable the generation of a reliable calibration model for qualitative or quantitative detection. In this thesis, filter design and wavelet transform methods are investigated as preprocessing tools to help extract the analyte signal from the underlying background. Numerical optimization is a key component of this work because of the need to select the appropriate filter or wavelet design parameters along with the related parameters that govern the calibration model. In this research, particle swarm optimization (PSO) is employed. Details about the signal processing and optimization calculations will be introduced in Chapter 3. The related regression methods for use in building quantitative calibration models will also be discussed in this chapter.

The work on data analysis methodology for use in process monitoring employed both laboratory simulations of continuous monitoring applications and actual process data. In Chapter 5, we used a four-component flowing system to investigate the coupling of PSO, digital filtering, and partial least-squares (PLS) regression. Both short- and long-term prediction data were studied. The long-term study was valuable in testing the stability and reliability of the calibration. To enhance prediction performance, a model updating approach was also investigated.

With the promising results from Chapter 5, Chapters 6 and 7 move forward to process monitoring of cell bioreactor runs based on the yeast, *Pichia pastoris*. The system is dynamic and complicated due to the cell growth. To obtain a real-time response, obtaining sufficient reference data for use in model building is impractical. The dynamic nature of the sample also dictates that a calibration model may not be transferable from run to run. Therefore, instead of PLS regression, a technique that requires an extensive set of calibration data, augmented classical least-squares (ACLS) modeling, was employed.

The ACLS method depends on knowledge of the chemical system under study in the form of either pure-component spectra of the sample constituents or composite spectra that describe the components of the spectral background. If these spectra are available, the ACLS method can be implemented without a large set of calibration data of known samples. Lack of reference data is still troublesome, however, because optimization requires a set of standard data with known analyte concentrations to drive the search process. A synthesis method is introduced in Chapter 6 to support the optimization of the calibration models for the bioreactor runs.

Finally, in Chapter 8, overall conclusions are drawn regarding the research results presented in Chapters 4 - 7. Suggestions are also provided for the direction of future research based on the current results and methods.

CHAPTER 2
INFRARED MEASUREMENT TECHNIQUES

This chapter provides an overview of the fundamental principles of Fourier transform-based infrared instrumentation and the associated raw data processing. A filter-based spectrometer is also introduced, followed by a review of the basic theory and radiance models involved in passive remote sensing.

2.1 Introduction to Infrared Spectroscopy

Infrared (IR) spectrometry is one of the most widely applied analysis methods for molecular species. The IR measurement focuses on a range of characteristic frequencies of light and the molecular absorption, emission or reflection phenomena that occur at those frequencies. The IR region generally covers from 780 nm to 1000 μm or in wavenumber, 12,800 to 10 cm^{-1}. Based on the energy and type of vibrational transitions, the IR region is further subdivided into near-, mid- and far- infrared regions.

The near-IR (NIR) region spans from 12,800 to 4000 cm^{-1} (780 nm to 2.5 μm). Spectral features in this region are composed of broad bands of overtones and combinations of fundamental vibrations of C-H, O-H and N-H bonds in organic functional groups. Initially, the NIR region was not frequently used because the spectral bands are weak and overlapped and the spectrum is complex for quantitative analysis. With the development of instrumentation and data analysis techniques, however, this region gained increasing interest due to its compatibility with aqueous samples. Nowadays, NIR spectroscopy has been successfully applied in agricultural, food, pharmaceutical, and petroleum industries, as well as in biological and environmental analysis.[21]

The mid-IR is the most widely used region in the overall IR spectrum. The mid-IR range covers between 4000 and 200 cm^{-1} in wavenumber or 2.5 to 50 μm in wavelength. This region involves the fundamental vibrations of most common chemical bonds. Because of the high specificity and relatively good sensitivity, it has been successfully used for identification of organic functional groups and consequently quanlitative and quantitative analysis of organic compounds. In this thesis, the mid-IR region is used for the quantitative analysis of gas phase organic compounds (e.g., methanol and ethanol) in passive FT-IR remote sensing. This will be introduced in Chapter 4.

Compared to the near- and mid-IR regions, the far-IR, ranging from 200 to 10 cm^{-1} (50 to 1000 μm) has historically been limited in application due to a lack of good light sources and detectors. The region can also provide unique spectral features in lower frequency, as the fundamental vibrations of organometallic and inorganic molecules (e.g., metal-ligand compounds) are characterized by heavy atoms and weak bonds. Additionally, the lattice vibrations of crystalline materials and electron valence or conduction band transitions in semiconductors also fall in this region.[22]

2.2 Fourier Transform Infrared Spectrometry and Raw Data Processing

2.2.1 Michelson Interferometer

The key part of a Fourier transform infrared (FT-IR) spectrometer is the optical component termed an interferometer. Among different types of interferometers, the double-beam interferometer designed by Michelson in 1891 is the most widely used in commercial FT-IR spectrometers. Next, an introduction to the Michelson interferometer is provided.[23,24]

Figure 2.1 depicts a schematic diagram of a Michelson interferometer. The radiation from the IR source is divided into two parts by a beamsplitter. One part of the beam is reflected to the stationary mirror, while the other part transmits through the beamsplitter to the movable mirror. Ideally, the beamsplitter should reflect half and transmit half of the incident beam. The divided beams recombine after reflection back to the beamsplitter by the mirrors. Again, the recombined ray is divided, and part transmits and part reflects. One beam passes through the sample and reaches the detector, and the other heads back to the light source. The beamsplitter plays an important role in dividing the incident light. The coating materials on the beamsplitter should have a high refractive index. The commonly used materials can be KBr, quartz, ZnSe, ZnS, CaF_2, depending on the specific spectral region to be investigated.

With the arrangement of the movable and stationary mirrors, a constructive or destructive interference between the reflected waves occurs due to the difference in travel distances. The optical path difference governed by the relative mirror positions is termed the retardation, δ. If the two waves are in phase, where $\delta = n\lambda, n =$

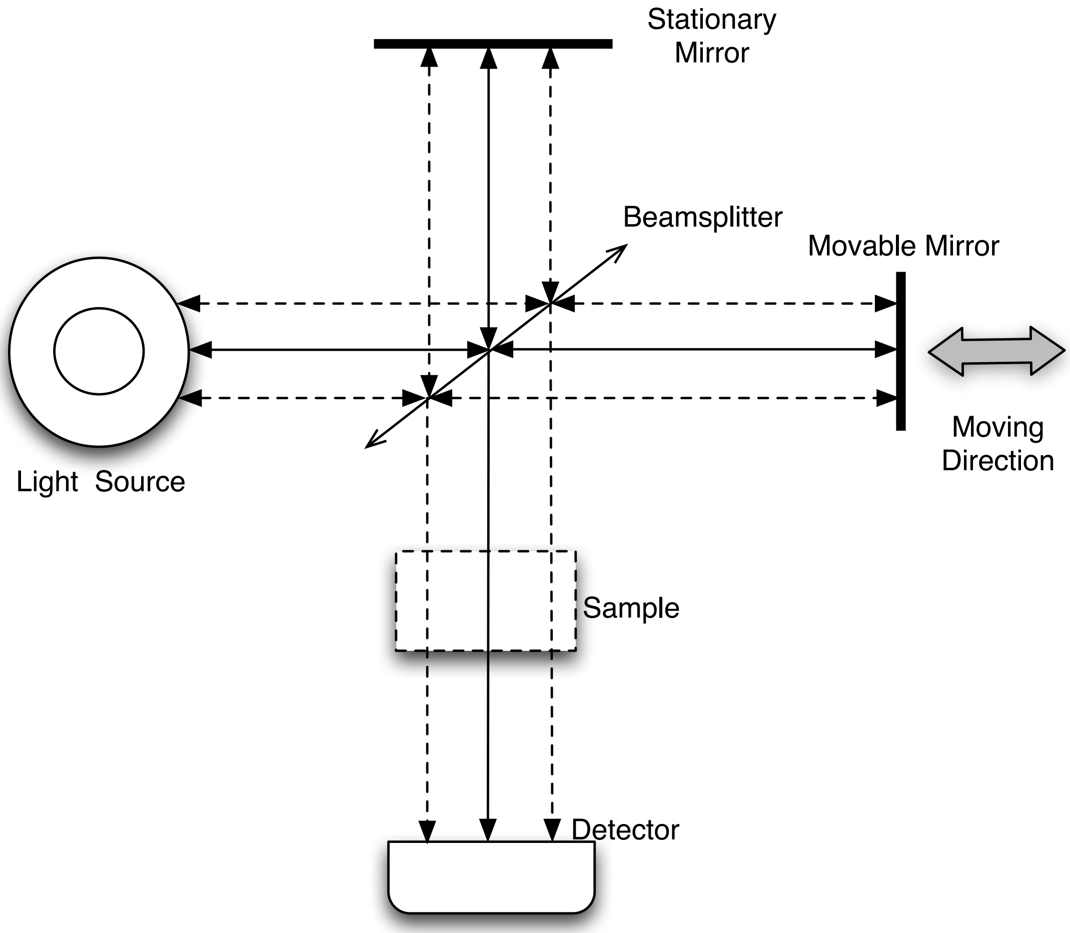


Figure 2.1. Schematic of a Michelson interferometer as used in a laboratory FT spectrometer. Radiance from the light source is directed to the beamsplitter, producing separate light beams that are directed to the fixed and movable mirrors. After reflection, the two beams recombine at the beamsplitter and undergo interference on the basis of the difference in paths traveled. The resulting light beam is directed through the sample and onto the detector. The detector signal is recorded as a function of the position of the moving mirror, resulting in an interferogram. In the remote sensing research presented in Chapter 4, the sample lies outside of the spectrometer and the light source is replaced by a set of entrance optics that direct an external source of radiance into the interferometer.

$0, 1, 2, ...$, constructive interference occurs, and a signal twice the amplitude of the original split beams is obtained (i.e., at an amplitude equal to the original light beam incident on the beamsplitter). For one-half wavelength out of phase ($\delta = n\lambda/2, n = 1, 3, 5, ...$), the two beams destructively interfere and cancel out. For phase differences in between the two extremes, the beams undergo partial interference. A plot of light output versus retardation is called an interferogram. For a monochromatic light source, the interferogram can be represented by a cosine wave as[5]:

$$I(\delta) = B(\bar{\nu})\cos(\frac{2\pi\delta}{\lambda}) = B(\bar{\nu})\cos(2\pi\bar{\nu}\delta) \tag{2.1}$$

where $I(\delta)$, a function of the retardation, δ, is the light intensity reaching the detector, and $B(\bar{\nu})$ is a constant corresponding to the light source, beamsplitter efficiency and detector response. This term is a function of the wavenumber ($\bar{\nu}$) of the light, which equals to $1/\lambda$, where λ is the wavelength of the light.

For a broad band continuous or a polychromatic light source, Eq. 2.1 can be extended to

$$I(\delta) = \int_{-\infty}^{+\infty} B(\bar{\nu})\cos(2\pi\bar{\nu}\delta)\,\mathrm{d}\bar{\nu} \tag{2.2}$$

The location where $\delta = 0$ is called the point of zero path difference (ZPD). At ZPD, all the light frequencies interfere constructively. As a consequence, the interferogram has its maximum intensity at this point. This point is also termed the centerburst. As the movable mirror moves away from the centerburst, the interferogram intensity decays. The wider the bandwidth of the incident light, the faster the decay. Therefore, an infinitely broadband light source will generate an interferogram with only the centerburst. For a perfect monochromatic light source, an infinite interferogram can be obtained without any decay.

2.2.2 Fourier Transform

To convert the interferogram from the time domain of the measured cosine signals to the frequency domain of a spectrum, Fourier transform (FT) is performed on the interferogram by

$$B(\bar{\nu}) = \int_{-\infty}^{+\infty} I(\delta)\cos(2\pi\bar{\nu}\delta)\,\mathrm{d}\delta \tag{2.3}$$

Because $I(\delta)$ is symmetric about the ZPD, Eq. 2.3 can be rewritten as:

$$B(\bar{\nu}) = 2\int_{0}^{+\infty} I(\delta)\cos(2\pi\bar{\nu}\delta)\,\mathrm{d}\delta \tag{2.4}$$

Equations 2.2 and 2.3 are called a Fourier transform pair, which demonstrate the relationship between the spectral domain single-beam intensity and the interferogram intensity.

2.2.3 Resolution and Sampling

The relationship in Eq. 2.4 indicates that the mirror must move towards infinity to obtain an infinite retardation. In practice, however, the movable mirror can only move for a certain distance. The FT can only apply from 0 to the maximum retardation in a finite interval. Therefore, the transformed single-beam spectrum has a finite resolution. The resolution or the difference between two neighboring frequencies ($\Delta\bar{\nu}$) is inversely proportional to the maximum scanned retardation as $\Delta\bar{\nu} = 1/\delta_{\max}$. To double the resolution, the mirror needs to move twice the original distance.

A relationship also exists between the spectral resolution and the signal-to-noise ratio of the measurement. Given that the measurement of each interferogram point contains both signal and noise, increasing the retardation will increase the total noise of the interferogram. Given that the spectrum has a finite bandwidth (range of frequencies), more total noise in the interferogram will lead to more total noise across the fixed bandwidth. Doubling the length of the interferogram will typically increase the noise level by a factor of $2^{1/2}$.

According to the Nyquist sampling theorem, a continuous signal must be sampled at a minimum rate of twice the maximum frequency of the signal. A lower sampling rate can induce a lower apparent frequency than the true frequency in the signal. This phenomenon is called aliasing. For a given resolution, $\Delta\bar{\nu}$, and maximum frequency, $\bar{\nu}_{\max}$, the required number of points in the interferogram, N_s is

$$N_s = \frac{2\bar{\nu}_{\max}}{\Delta\bar{\nu}} \tag{2.5}$$

In the remote sensing project described in Chapter 4, interferograms were collected under the control of the modulated frequency of a reference He-Ne laser with a frequency of 15798 cm^{-1}. The laser was directed through the interferometer and onto a dedicated detector. The signal from the laser detector was input to a sampling circuit which coordinated the acquisition of data from the infrared detector. Interferograms were sampled at every eight zero-crossings of the laser interferogram, which dictated the maximum frequency according to the requirements of the Nyquist theorem. Therefore, light frequecies up to 1975 cm^{-1} (1/8 of the maximum laser frequency) could be sampled without aliasing.

The choice of sampling rate is dictated by the maximum frequency of the light source or maximum response frequency of the detector. The choice of 1975 cm^{-1} was driven by the use of ambient infrared light as the light source in the remote sensing measurements described in Chapter 4. Essentially no light above 1975 cm^{-1} is observed from natural sources, and the Hg:Cd:Te (MCT) detector used in this work had its response range restricted to this wavenumber maximum.

2.2.4 Phase Correction

Theoretically, the collected interferogram should be symmetric with respect to the ZPD. However, in reality, this case is not guaranteed. The asymmetry is primarily caused by the phase shift introduced by reflections within the beamsplitter. Mathematically, this causes the interferogram represented by Eq. 2.2 to contain both cosine and sine components. Therefore, to correct the shift and restore a symmetric interferogram, a phase correction process is needed.

Phase correction can be applied to both the interferogram and single-beam spectrum. In this thesis, the Mertz method was performed in the spectral domain.[25–27]

In this method, the cosine and sine components are separated by real and imaginary parts in the FT, respectively. The complex FT can be written as:

$$B(\bar{\nu}) = \mathrm{Re}(\bar{\nu})\cos\theta(\bar{\nu}) + \mathrm{Im}(\bar{\nu})\sin\theta(\bar{\nu}) \tag{2.6}$$

where Re and Im represent the real and imaginary parts in $B(\bar{\nu})$. The term, $\theta(\bar{\nu})$, is the wavenumber-dependent phase error function, defined as:

$$\theta(\bar{\nu}) = \tan^{-1}\left(\frac{\mathrm{Im}(\bar{\nu})}{\mathrm{Re}(\bar{\nu})}\right) \tag{2.7}$$

2.2.5 Apodization

As mentioned above, the interferogram can only be collected in a limited retardation in practice. However, the FT is applied from 0 to infinity as in Eq. 2.4. Therefore, the FT is not applied to the true interferogram, but rather to a signal equal to the product of the true interferogram and a boxcar sampling function. Given that the maximum retardation is Δ, the boxcar truncation function can be written as:

$$D(\delta) = \begin{cases} 1 & \text{if } -\Delta \leq \delta \leq +\Delta \\ 0 & \text{if } \quad \delta > |\Delta| \end{cases} \tag{2.8}$$

With the truncation sampling function, Eq. 2.4 can be rewritten as:

$$B(\bar{\nu}) = 2\int_{0}^{+\infty} I(\delta)D(\delta)\cos(2\pi\bar{\nu}\delta)\,\mathrm{d}\delta \tag{2.9}$$

Because multiplication of two functions results in a convolution calculation in the FT, the infinite interferogram function is convolved with the FT of the boxcar truncation function (sinc function of $\sin x/x$). If the original spectral signal is monochromatic, an infinitely narrow line can be observed after the FT if the true interferogram has been acquired. However, with the sinc function determining the characteristics of the spectral band and shape, the resulting FT signal will have artifacts such as oscillating side-lobes. In this case, the width of a spectral band is the spectral resolution associated with the measurement. For the boxcar function, the width at half-maximum height of the observed band is $1.207/\Delta$.

To reduce the artifacts caused by the boxcar function and effectively sample the interferogram, a triangular-shaped or other sampling function ($A(\delta)$) is often applied to replace the boxcar truncation function. Rather than sharply transitioning to zero on the edges, these functions gradually taper to zero, resulting in fewer artifacts when transformed to the spectral domain. This procedure of multiplying the interferogram by an artificial sampling function is termed apodization. The triangular apodization function can be represented as:

$$A(\delta) = \begin{cases} 1 - |\delta/\Delta| & \text{if} - \Delta \leq \delta \leq +\Delta \\ 0 & \text{if} \quad \delta > |\Delta| \end{cases} \tag{2.10}$$

The corresponding FT of the apodized interferogram is

$$B(\bar{\nu}) = 2 \int_0^{+\infty} I(\delta) A(\delta) \cos(2\pi\bar{\nu}\delta)\, \mathrm{d}\delta \tag{2.11}$$

As noted above, compared to the boxcar function, other apodization functions can minimize the side-lobe artifacts caused by the boxcar truncation. A price must be paid in resolution, however. The resolution with the triangular sampling function will increase to $1.772/\Delta$ while the oscillating artifacts propagating away from the true spectral frequency are reduced by up to a factor of 5. In the FT calculation in Chapter 4, triangular apodization was employed.

2.2.6 Advantages of FT-IR Spectrometry

In contrast to traditional dispersive instruments, FT-IR spectrometry offers several advantages. The multiplex advantage (Fellgett advantage)[5,28,29] arises from the fact that each sampling of the IR detector (i.e., each interferogram point) contains information from all wavelengths in the input light. This qualifies as a multiplex measurement, defined as a measurement process in which multiple channels of information are carried on a single measurement channel. By contrast, in a conventional dispersive instrument, one measurement point includes information from only one spectral resolution element. Thus, for a fixed number of samples of the IR detector, the resolution elements are sampled more often in the FT-IR measurement. Relative to the dispersive instrument, this translates to either a greater SNR in the same measurement time or the same SNR in less measurement time. This fast-scanning capability facilitates the use of signal averaging to reduce the level of random noise in the spectrum.

The second advantage is termed the Jacquinot advantage,[5,28] which is related to the high light throughput in the FT-IR instrument. As no slits are used in the optical path, more light can reach the detector relative to dispersive instruments. If the measurement is detector-noise limited (i.e., the noise level of the measurement is driven by the intrinsic noise of the detector), higher light throughput will lead to higher incident power onto the detector and thus, a higher SNR. This advantage is limited to some extent since too much incident light may saturate the detector or cause it to respond nonlinearly. Consequently, for a relatively transparent sample, the light may need to be attenuated before reaching the detector. However, high throughput is especially valuable for highly absorbing or scattering samples or in light-limited measurements such as passive IR remote sensing.

The third advantage is the highly repeatable spectral wavelengths, called Connes advantage.[5,30] With the reference He-Ne laser providing the control of the sampling of the interferogram, the spectral frequency is directly related to the stability of the reference laser.

Meanwhile, FT-IR spectrometry also has potential disadvantages. The setup of the movable mirror in collecting the interferogram requires a high-precision driving and alignment system, thereby affecting the ruggedness and portability of the spectrometer. A second limitation is termed the “multiplex disadvantage”.[29] The multiplex nature of the FT-IR measurement dictates that if a given resolution element has a high noise value, this increased noise will “contaminate” each interferogram point and ultimately be distributed across the entire spectral bandwidth. In this thesis, the passive remote sensing project employed the FT-IR spectrometer. In the

NIR monitoring chapters, a filter-based spectrometer experimental set up was investigated instead. Next, the filter based spectrometer (acousto-optic tunable filter) is introduced.

2.3 Acousto-Optic Tunable Filter

The acousto-optic tunable filter (AOTF) is an all-solid state electronically tunable spectral band-pass filter.[6,31,32] The principal theory of the AOTF has been established since the 1920s and the acousto-optic phenomenon was observed in the 1930s for the first time. The AOTF is constructed from a specially cut anisotropic crystal bonded with an array of acoustic transducers. The anisotropic crystal is typically quartz or tellurium dioxide (TeO_2). The acoustic transducer is made of a piezoelectric material, by which acoustic waves are generated and launched into the crystal when a radio-frequency (RF) electrical signal is applied into the transducer. As the acoustic waves propagate through the crystal, a periodic moving grating is produced to diffract portions of the incident beam. Figure 2.2 shows a schematic of the operation of an AOTF.[33]

The AOTF is compatible in selecting both single wavelength and multiple wavelengths from the incoming light and choosing from different sources, for example multi-line sources such as lasers or broadband light sources such as tungsten-halogen lamps. After a non-polarized light beam passes the crystal, the light beam is split into three beams, the undiffracted beam, the ordinary beam and the extraordinary beam. The ordinary and extraordinary beams, which are generated by the acousto-optic effect, are orthogonally polarized. If the incident light is an extraordinary ray, it will be converted into an ordinary ray and spatially separated from the original extraordinary beam by interaction with the acoustic wave propagating in the AOTF. Since the polarized and the original beam are different in polarization, the corresponding refractive indices are different. For a fixed acoustic frequency, only a narrow band of optical frequencies can be diffracted for analytical purposes. Therefore, the spectral bandpass can be tuned over a large optical region by simply changing the frequency of the applied radio frequency.

A brief introduction of the working theory of the AOTF is discussed below to help understand the diffraction process in the crystal.[6,31,33–35] The diffraction can be considered as a transfer of energy and momentum of the electromagnetic wave, and the energy and the momentum should be conserved. The following equation shows the conservation of the momentum:

$$\mathbf{k}_d = \mathbf{k}_i \pm \mathbf{k}_s \tag{2.12}$$

The terms $\mathbf{k}_d$, $\mathbf{k}_i$, and $\mathbf{k}_s$ in Eq. 2.12 represent the wave vectors of the diffracted light, the incident light and the acoustic wave, respectively. The momentum of the incident and diffracted light can be written as:

$$|\mathbf{k}_d| = \frac{2\pi n_d}{\lambda} \tag{2.13}$$

$$|\mathbf{k}_i| = \frac{2\pi n_i}{\lambda} \tag{2.14}$$

where n_d and n_i are the refractive indices of the diffracted and incident light, respectively, and λ is the wavelength of light diffracted by the acoustic wave.

If the AOTF is collinear, shown in Figure 2.3, the incident light, the diffracted light and the acoustic wave are all collinear. As mentioned above, if the incident light is an extraordinary beam, the diffracted light is an ordinary beam. Then the conversion of the momentum can be written as

$$\mathbf{k}_d = \mathbf{k}_i - \mathbf{k}_s \tag{2.15}$$

$$|\mathbf{k}_s| = \frac{2\pi f_s}{v_s} \tag{2.16}$$

$$\lambda = \frac{v_s(n_e - n_o)}{f_s} \tag{2.17}$$

In Eq. 2.16, the f_s and v_s are the frequency and velocity of the sound wave, respectively.

For a non-collinear AOTF, the incident light, the diffracted light and the acoustic wave are not collinear. The wavelength of light diffracted by the sound wave is

$$\lambda = \frac{v_s(n_e - n_o)(sin^2\theta_i - sin^4 2\theta_i)^{1/2}}{f_s} \tag{2.18}$$

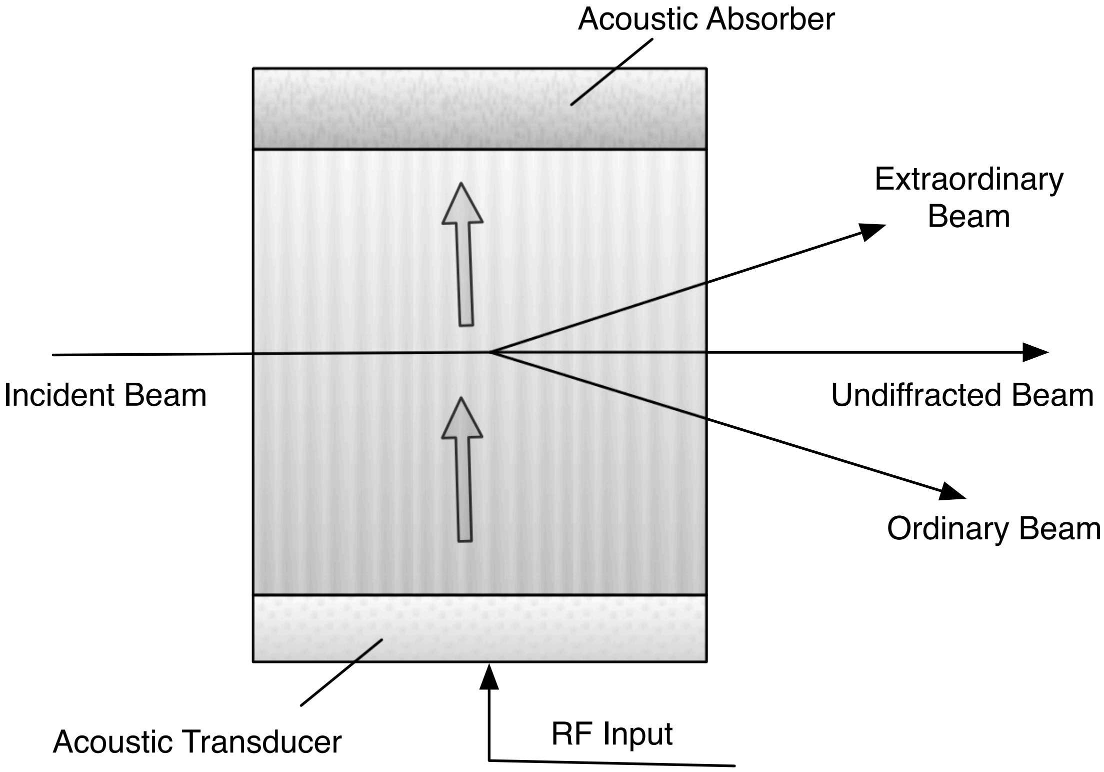


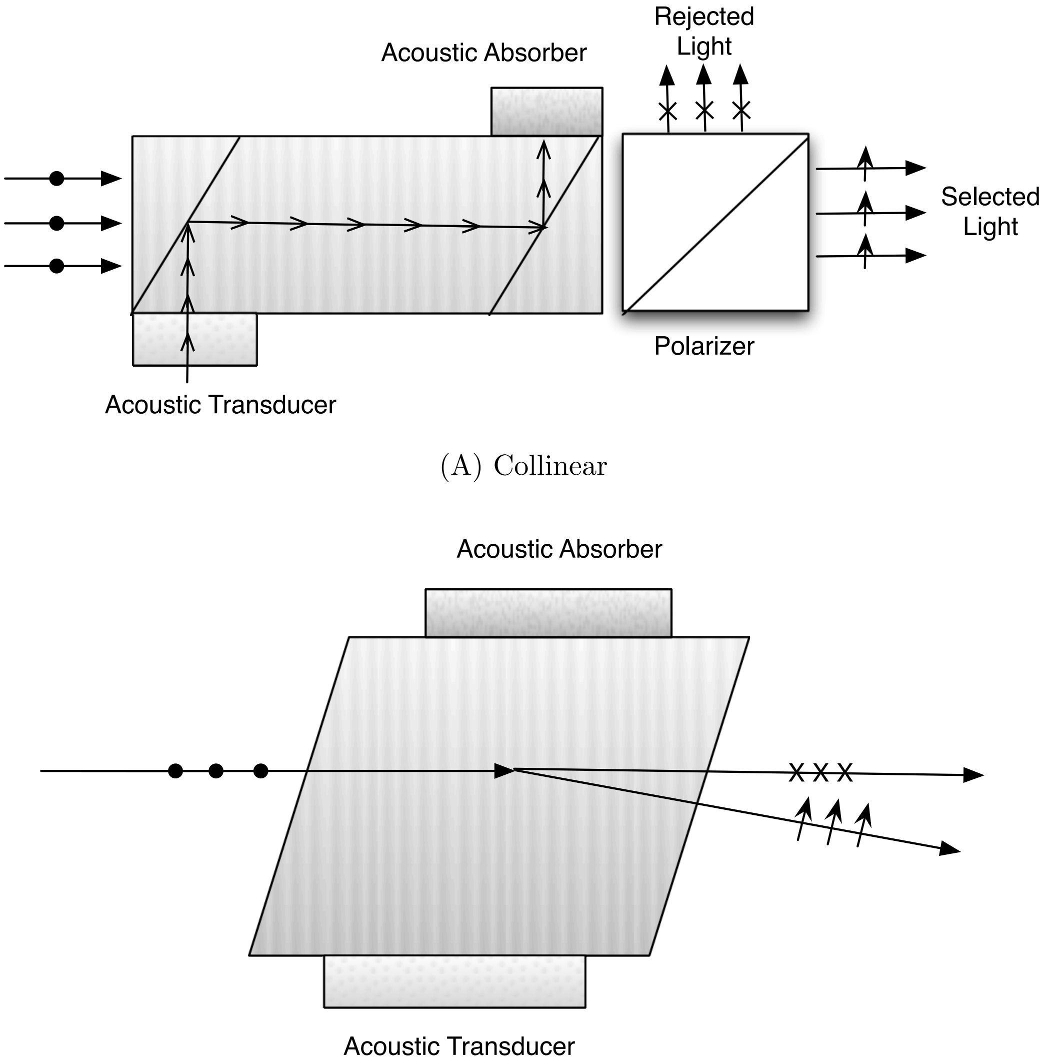
Figure 2.2. Schematic diagram of the operation of an AOTF. An RF signal is launched into the crystal, inducing a pressure wave that causes a change in refractive index. The beam is diffracted according to a relationship between the wavelength of the incident light and the frequency of the RF energy.

Here the θ_i is angle of the incident light. If the AOTF is all collinear, the incident angle is equal to 90°, then Eq. 2.18 will reduce to Eq. 2.17.

From the equations above, given the frequency of the acoustic wave, only certain wavelengths of light can be diffracted from the crystal. Therefore, by changing the electrical signal of the RF applied to the acoustic transducer, the frequency of the acoustic wave passed through the crystal is altered, and a specific spectrum can be achieved by the diffraction.

Based on the optical configuration and acoustic wave vectors, there are two different types of AOTFs, collinear and noncollinear.[6] For the collinear AOTF in Figure 2.3A, the acoustic wave and the incident beam propagate the crystal collinearly during the acousto-optic interaction, as does the diffracted beam whose polarization is orthogonal to that of the incident beam. The undiffracted (zero-order diffraction) and diffracted beams also emerge collinearly from the AOTF. Since the polarization of the diffracted beam is different, it can be separated from other beams by a polarizer. Since the interaction length between the incident beam and acoustic wave is relatively long in this type of filter, crystals that have smaller acoustic figures of merit (e.g. quartz and MgF_2) are often used for collinear type of AOTFs.

However, the noncollinear AOTF is more common since the construction of the collinear device is sometimes impossible due to the crystal structure. Figure 2.3B shows the case when the incident, diffracted and acoustic waves are noncollinear. The acoustic wave diffracts the vertically polarized incident light into a horizontally polarized beam. The diffracted light can be readily isolated because the transmitted



(A) Collinear

(B) Non-collinear

Figure 2.3. Two types of AOTF designs. The non-collinear design is the most common configuration.

beam is well separated from the diffracted beam. If the incident light is horizontally polarized, it will be diffracted into a vertically polarized beam. For the non-polarized incident beam, it will be diffracted into two orthogonally polarized beams, which propagate in different directions. Designs of this type are constructed from materials with high acousto-optic figures of merit, such as TeO_2.

Hence, the AOTF is similar to the diffraction grating. The grating constant in this case is the frequency of the acoustic wave introduced into the crystal. One of the advantages of the AOTF is the frequency of the acoustic wave can be electronically selected. The wavelength scan rate could be very fast, since the transit time of the acoustic wave across an optical beam is on the order of microseconds. This translates to a fast tuning speed of the filter.

Another advantage of the AOTF is the flexibility in the wavelength selection. The spectral range can be widely selected from the ultraviolet through the visible to the infrared. The wavelength of diffracted light can be specified by tuning the RF applied to the crystal (Eqs. 2.17 and 2.18). It also allows multiple wavelengths of diffraction if more than one RF signal is applied at the same time, which makes it possible to avoid scanning all wavelengths as in the conventional FT-IR spectrometer. The AOTF used in this research is a polychromator to provide multiple wavelengths simultaneously. The width of the filter bandpass is 24 cm^{-1}.

Compared to the conventional FT-IR spectrometer, the AOTF is more rugged since there is no moving parts. It offers comparable high light throughput with the FT-IR spectrometer. However, the AOTF has not been widely commercially applied in NIR applications because there are still engineering problems that have not been fully solved and this technology is not technically mature. Hence, it is still under development.[31,34]

Other main components used in the AOTF instrument are similar to those used with an FT-IR spectrometer. The NIR light source was a tungsten-halogen lamp. The detector used in this research was a thermoelectrically cooled InGaAs semiconductor detector.

2.4 Passive Infrared Remote Sensing

Passive infrared remote sensing is a measurement employed with an emission spectrometer in which a naturally occurring source of IR radiance serves as the light source. Because natural sources are typically at ambient temperature, the radiant powers are weak, and the measurement platform is typically a high-throughput FT-IR spectrometer.[36,37]

Figure 2.4 depicts the typical experimental setup of the passive measurement. A target vapor lies within the field-of-view (FOV) of the spectrometer and is observed against a background scene. Telescope entrance optics are typically employed with the spectrometer to restrict the FOV and thus allow the measurement to focus on a specific spatial location. The spectrometer collects the IR radiation emitted from the background, any emission of light from the target vapor cloud, any atmospheric emission, and the self-emission from the spectrometer itself.[37–39]

As a function of wavelength, λ, the radiance ($L(\lambda)$) emitted by any surface can

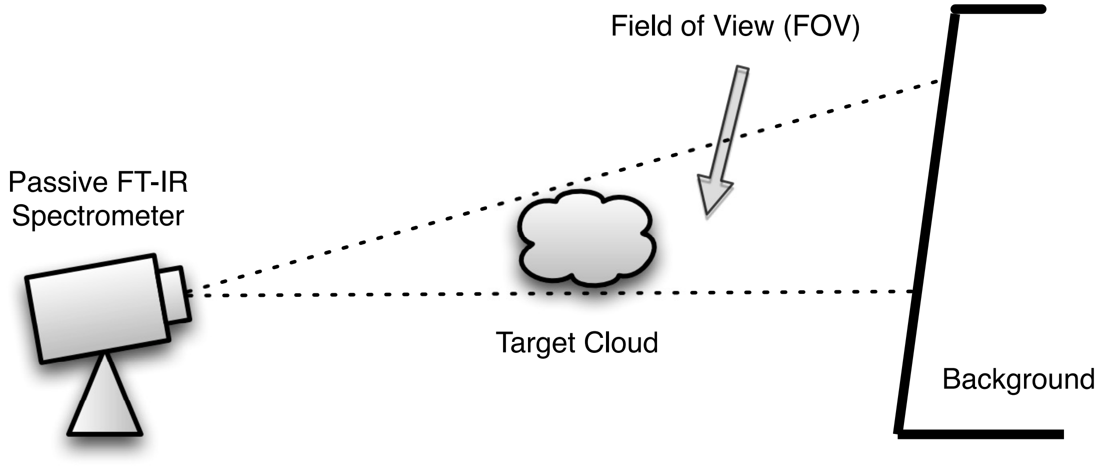


Figure 2.4. Schematic of a passive FT-IR measurement. The sample lies in the intervening atmosphere between an emission FT-IR spectrometer and a naturally occurring IR background. Telescope entrance optics are typically placed on the spectrometer to restrict its FOV.

be scaled by the radiance from a perfect Planck's blackbody which is solely dependent on the temperature of the material:

$$L(\lambda) = \epsilon(\lambda) \times L^*(\lambda, T) \tag{2.19}$$

Here, $L^*(\lambda, T)$ is the theoretical Planck's blackbody function and $\epsilon(\lambda)$ is the emissivity term, which is defined by the ratio of the radiance from the given body and the perfect blackbody ($\epsilon = 1$). Usually, the emissivity is wavelength-dependent and a material with this property is called a grey body.

The radiance of the gas can be transmitted, absorbed or reflected. The total power law indicates that the sum of transmittance (τ), emittance(ϵ) and reflectance(ρ) equals to unity. With the assumption of no reflectance for gas samples, the emissivity can be calculated from the transmittance term by $\epsilon(\lambda) = 1 - \tau(\lambda)$.

As mentioned above, the total IR radiance incident on the sensor is the sum of radiances from the attenuated parallel layers from the background, target gas cloud and the intervening atmospheric gases. The total spectral radiance (L_x) can be expressed as:

$$L_x = \tau_t \tau_a L_{bkg} + (1 - \tau_t \tau_a) L_t \tag{2.20}$$

where τ_t and τ_a denote the transmittance of the target gas cloud and the atmosphere, respectively. The terms, L_{bkg} and L_t, are the radiance values of the background and the target cloud, respectively. Each term in Eq. 2.20 is wavelength dependent. The first term in Eq. 2.20 can be considered an absorption term where radiance emitted from the background is attenuated by the atmosphere and target gas. The second term can be considered an emission term that describes direct IR emission from the target gas. Note also that the radiance model in Eq. 2.20 assumes there is no direct radiance from the atmosphere itself. If this assumption were not made, Eq. 2.20 would be extended by adding a radiance contribution from the atmosphere.

Analyte information enters the model through the $\tau(t)$ term, as transmittance is defined as

$$\tau_t(\lambda) = \exp(-\alpha(\lambda)cl) \tag{2.21}$$

where $\alpha(\lambda)$ is the absorptivity of the analyte molecule at wavelength, λ, c is the concentration, and l is the depth of the cloud along the optical path.

The key requirement of the radiance model is that in order for analyte information to be obtained in the passive IR measurement, a significant difference must exist between the temperature of the target cloud (T_t) and the radiant temperature of the IR background (T_{bkg}). Assuming the background and target cloud are perfect blackbodies, both radiance terms in Eq. 2.20 can be represented by the theoretical Planck's function values. Accordingly, their radiance is only dependent on temperature. In this case, if the temperature difference between them is small, for example $T_{bkg} = T_t$ at the extreme case, $L_{bkg} = L_t$. Then, Eq. 2.20 can be reduced to $L_x = L_{bkg} = L_t$. No analyte information is obtained in this case. Essentially, the rates of absorption and emission of the analyte are equal.

The measurement thus becomes challenging as T_{bkg} approaches T_t. Another challenging case occurs when the concentration of the detected gas is low, or the cloud depth is small. In these cases, the value of τ_t is close to unity and the transmittance term has little impact on the measured radiance. A final consideration is that in passive remote sensing, depending on the temperature relationship noted above , the observed IR spectral features of the target gas could be either emission ($T_t > T_{bkg}$) or absorption ($T_t < T_{bkg}$). Applications of this radiance model in gas detection will be discussed in Chapter 4.

CHAPTER 3
SIGNAL PROCESSING AND DATA ANALYSIS

In IR spectroscopy, data preprocessing is a necessary procedure because of the potential instability of the spectral background, interference from non-analyte species present in the sample, and the occurrence of measurement noise and spectrometer drift. The relatively weak absorptivities in the infrared region dictate that the spectral signal-to-noise ratio (SNR) will always be a concern and that spectral artifacts such as baseline variation will be more prominent than in other spectral regions were the analyte signals are larger.

Signal processing methods such as digital filtering and wavelet analysis are frequently used to eliminate extraneous spectral signals, thereby helping to uncover useful information related to the target analyte under study. In this chapter, a general review of signal processing techniques is provided.

In the quantitative analysis of IR spectra, after the data preprocessing step, a calibration model is built to allow the estimation of analyte concentration. Because of the common occurrence of spectral overlap, few real-world multicomponent samples can be quantified by use of measurements at a single wavelength. Thus, full spectra are collected and usually a multivariate statistical regression is performed to build the calibration model. The regression model can be based on partial least-squares (PLS), classical least-squares (CLS), or a variety of other related methods. The mathematics and statistics related to these regression methods will also be introduced in this chapter.

Additionally, in both signal processing and multivariate regression, parameter optimization is always required to help select the optimal parameter combinations that define the specifics of the method. In this study, particle swarm optimization (PSO) was investigated for its use in controlling and optimizing parameter selection. The optimization procedure of this method will also be discussed in this chapter.

3.1 Signal Processing

For all instrumental measurements, interference signals commonly exist and are superimposed on the analyte signal. These interference signals include signals from instrumental noise, baseline drift, and intensity changes due to variation in the light source or the detector response. It is desirable to eliminate these unwanted signals before further quantitative analysis is performed.

Instrumental noise can be reduced by spectral co-addition or using a smoothing digital filter. Adding a linear or nonlinear trend in the model can help correct the baseline drift. These effects can also be minimized through the computation of absorbance spectra, based on taking the ratio of sample and background spectra. This procedure is effective only if the background spectrum carries the same information as in the sample spectrum regarding the artifact to be removed. Only in this case will the artifact be removed through the spectral ratio. In this section, two data preprocessing techniques used in this dissertation, digital filtering and wavelet analysis, will be introduced.

3.1.1 Digital Filtering

The principle of digital filtering is to separate components of a measured signal on the basis of the underlying frequency components that comprise it. Here, the term, frequency component, does not refer to a spectral frequency, but rather to the frequency of a sine or cosine waveform used to model the input data. When applied to IR data, the typical purpose of digital filtering is to remove unwanted signals and thereby help to extract the information regarding the analyte of interest. The implementation of this approach is based on the assumption that the analyte and the non-analyte information can be decomposed into their underlying harmonic frequencies.

Depending on the selected range of signal frequencies, the digital filter design can be lowpass, highpass, bandpass or bandstop.[40–42] Digital filtering can be performed on both the interferogram obtained from an FT-IR measurement as well as the corresponding spectrum.[43–46] In this dissertation, digital filtering in the spectral domain was investigated.

As noted above, the spectral signal can be considered as a series of sine and cosine waveforms with varied phases and frequencies superimposed on each other. Lower sine and cosine frequencies model components of the spectrum that vary slowly. Normally, this frequency region is dominated by background information, baseline variation, or instrumental drift effects. Higher frequency sines and cosines represent rapidly changing spectral features, for example fast-varying random noise. Therefore, a lowpass filter which passes low frequencies while attenuating higher frequencies could suppress random noise and a highpass filter could remove broad features such as baseline variation by attenuating the lower frequency information while passing information at high frequencies. A bandpass filter could potentially extract specific analyte frequency components while eliminating both baseline variation and spectral noise. A bandstop filter can be used to suppress the signal at a specific frequency while passing all other frequencies.. This type of filter is similar to a notch filter usually used in Raman spectrometry.

The profile of the action of a filter on the basis of frequency is termed its frequency response function. A frequency response function plots a measure of the transmission or attenuation of the filter as a function of frequency. The frequency scale is typically presented in a dimensionless form, varying from 0.0 to 1.0, where 1.0 defines the maximum harmonic frequency of the data as determined by the sampling rate and the Nyquist theorem (see section 2.3 in Chapter 2).

The frequency response is divided into groups of three regions, termed passbands, stopbands, and transition bands. A passband defines a range of frequencies the filter will pass, while a stopband specifies a range of frequencies that will be suppressed. The transition band lies between the passband and stopband. It specifies a region in which the frequency response is allowed to undergo a transition between the passband and stopband. In this range of frequencies, partial suppression of the signal will occur. Lowpass and highpass filters have one passband, one stopband, and one transition band. A bandpass filter will typically have a single passband with stopbands and transition bands on each side.

Digital filter design begins with the specification of the target frequency response. This can be non-trivial because the spectral features, analyte or non-analyte, are not distributed clearly into different frequency ranges. Therefore, it is difficult to draw a sharp line between them. Attenuating an unwanted signal will always cause some loss of the analyte signal.

Because the frequency response pertains to the underlying harmonic components of the input data, it cannot be used directly to accomplish the filtering step. From the standpoint of the Fourier transform (FT), the frequency response exists in the frequency domain, while the input spectrum exists in the time domain. In this regard, the spectrum is envisioned as a signal that is sampled vs. time. To operate on the spectrum, a representation of the frequency response in the time domain must be obtained. This is termed the impulse response of the filter. Given an input frequency response function, digital filter design methods estimate the corresponding impulse response. The filter is then applied by computing the convolution of the impulse response with the measured spectrum. The output of this calculation is a filtered spectrum.

When employing digital data, the input signal is composed of discrete points that have been sampled. Therefore, in using the filter, the signal is operated on point-by-point from beginning to end to approximate the convolution of the data with the impulse response of the filter. Depending on how the convolution is estimated, digital filters are categorized into finite impulse response (FIR) and infinite impulse response (IIR) filters.[40,47] The factor that differentiates these two types of filters is how the feedback of the previous point in the data sequence is manipulated. An FIR filter only uses the point from the input signal, while the IIR filter includes the filtered output of previous points when calculating the filtered value corresponding to the current point. The different operations can be represented by the equations shown below:

$$y_n = a_0 x_n + a_1 x_{n-1} + a_2 x_{n-2} + ... + a_N x_{n-N} \tag{3.1}$$

$$\begin{aligned} y_n = a_0 x_n + a_1 x_{n-1} + a_2 x_{n-2} + ... + a_N x_{n-N} \\ -b_1 y_{n-1} - b_2 y_{n-2} - ... - b_M y_{n-M} \end{aligned} \tag{3.2}$$

The terms in Eqs. 3.1 and 3.2 show how the FIR and IIR filters treat an input signal. In these equations, y_n denotes the filtered discrete signal at point n and the x_i are points in the original input signal at point n and before. The values of a and b are the filter coefficients (i.e., points comprising the impulse response) obtained from the filter design. It can be seen in Eq. 3.1 that an FIR filter employee the current data point and the raw data points located before it. In Eq. 3.2, the IIR filter also includes the filtered signal from points prior to the current point.

The number of filter coefficients, N and M, in Eqs. 3.1 and 3.2 dictate what is termed the filter order. Specifically, the order is N for an FIR filter and the maximum of N and M for an IIR filter. As seen from an inspection of the equations, the filter order determines the computational demands of the filtering step. The purpose of filter design is to compute an impulse response whose action approximates the target frequency response as closely as possible while doing so with as low a filter order as possible.

In using an FIR design to obtain a narrow bandpass filter with a high degree of stopband attenuation, a high filter order is typically required, resulting in an inefficient calculation. The IIR design, because it takes advantage of the previously filtered points, generally needs fewer coefficients than FIR filters to achieve a similar performance. This results in computational savings when the filter is applied. An IIR filter will also typically allow a narrower transition band, thus resulting in a sharper rolloff between the passband and stopband.

To take advantage of the properties noted above, the digital filtering work described in this dissertation employed IIR filters. The most widely used IIR filters are based on the Butterworth, Chebyshev type I and II, and Elliptic design methods.[48,49] The Chebyshev type II design was used in this work. This filter design has a fast rolloff between the stopband and the passband which offers good flexibility in frequency selection.[50]

3.1.2 Wavelet Transform

Wavelet analysis is based on a similar transform method to the FT.[51,52] In the FT, a signal is decomposed into sine and cosine waves of different frequencies, while the transformation is between the time domain and the frequency domain. Fourier analysis is generally very useful because the frequency components of a signal are often very useful in characterizing it. One of the limitations of the FT, however, is that it is based on the assumption that the time base of all the frequency components is the same. Stated differently, the FT assumes that underlying harmonic frequencies are present throughout the sampled signal. There is no consideration that a given frequency may appear at some point after the sampling of the signal has begun. In effect, time information is lost, and one cannot obtain information about when a particular event occurs.

A key limitation of the FT is that the basis functions used in constructing the model for the input data (i.e., the sine and cosine waveforms) are infinitely long functions. By contrast, wavelet analysis is based on modeling an input signal with harmonic functions (i.e., wavelets) that are limited in time (i.e., have a starting and stopping point). The wavelet model is able to accomplish both time and frequency estimation. In the terminology of the field, the wavelet transform is able to provide a multi-resolution analysis of the input data.[53–56]

In the wavelet transform, a wavelet function is used to decompose an input signal by moving the wavelet window along the signal. Wavelet functions are more flexible in terms of shape and length than the sine and cosine functions used as the basis for the FT. To obtain precise low-frequency information, a longer window is needed. Correspondingly, a shorter window is desired if high-frequency components are of interest.

The wavelet transform can provide both global and detailed views of the input signal. This flexibility can be achieved by shifting the window in the time variable and dilation of the wavelet on the frequency variable by alteration of a template function called the “mother wavelet”. The mother wavelet describes a family of functions that are specified by the family name and an order parameter. Figure 3.1 displays wavelet functions in different orders from Symlet family.

A mother wavelet ($\Psi(\lambda)$) must belong to the absolutely squared integral function space L^2 and meet the admissibility condition as:

$$\int_{-\infty}^{+\infty} \frac{|\Im(\Psi(\lambda))|^2}{\omega} \mathrm{d}\omega < \infty \tag{3.3}$$

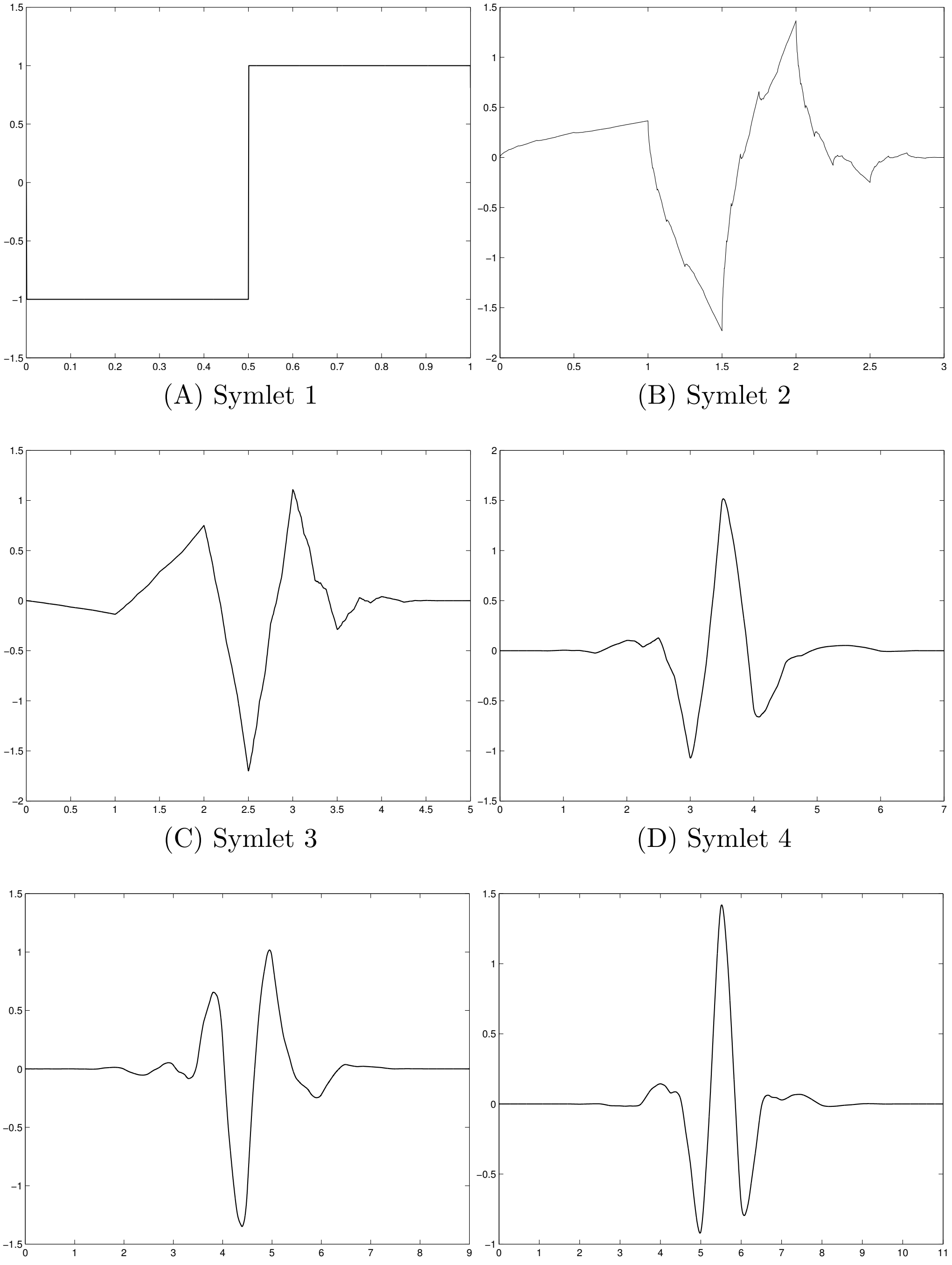
In the equation above, $\Im$ denotes the FT and ω is the radial frequency in the Fourier domain. A mother wavelet function must oscillate and have an average value of zero. It also needs to meet the requirement of exponential decay and be dually located in both time and frequency domains. Depending on the interested frequency region, a mother wavelet can be shifted or dilated by

$$\Psi_{a,b}(t) = \frac{1}{\sqrt{|a|}} \Psi(\frac{t-b}{a}) \tag{3.4}$$

In Eq. 3.4, a and b are the variables to control frequency dilation and time shifting, respectively. The term $1/\sqrt{|a|}$ is a normalizing constant to ensure the magnitude of the wavelet function is unity. Then, for a continuous signal $f(t)$, the wavelet transform is defined as:

$$W_f(a,b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{+\infty} \overline{\Psi(\frac{t-b}{a})} f(t)\,\mathrm{d}t \tag{3.5}$$

Similar to the FT, all signals used are digital data sampled at discrete points.



(A) Symlet 1

(B) Symlet 2

(C) Symlet 3

(D) Symlet 4

(E) Symlet 5

(F) Symlet 6

Figure 3.1. Wavelet functions with different orders from the Symlet mother wavelet. The subfigure caption shows the corresponding order of the function. The shapes of the function are different with different orders.

The discrete forms of Eqs. 3.4 and 3.5 are

$$\Psi_{m,n}(t) = a_0^{-m/2}\Psi(a_0^{-m}t - nb_0) \tag{3.6}$$

$$C_{m,n}(f) = \int_{-\infty}^{+\infty} \Psi_{m,n}(t)f(t)\,\mathrm{d}t \tag{3.7}$$

In Eq. 3.6, m and n are analogous to a and b in Eq. 3.4 and are used to scale the dilation and time shift. Usually, $a_0 = 2$ and $b_0 = 1$ in this equation and the function is called dyadic. But with Eq. 3.7, it is still difficult to calculate the wavelet transform. The discrete dyadic wavelet transform, in which $\Psi_{m,n}$ plays a role in the form of discrete filters, can decompose the given signal $f(t)$ into highpass and lowpass components through the impulse response functions, G and H. Functions G and H are the impulse responses of highpass and lowpass filters, respectively. The corresponding wavelet coefficients can be calculated by:

$$A_{2^{j+1}}f(n) = \sum_{l=0}^{L-1} H(l)A_{2^j}f(n - 2^j l) \tag{3.8}$$

$$D_{2^{j+1}}f(n) = \sum_{l=0}^{L-1} G(m)A_{2^j}f(n - 2^j m) \tag{3.9}$$

The terms, A and D, are termed the approximation and details coefficients, which represent the low- and high- pass parts respectively. To perform the discrete wavelet transform (DWT), the signal must be of length 2^n. Each decomposition step generates the approximation and details coefficients, whose lengths are equal to 1/2 of the decomposed signal. Symmetric extension or zero-filling can be used to extend the input signal to the nearest power of two.

When applied in signal processing, the wavelet decomposition can be performed iteratively for several steps on both the obtained approximation and details components depending on the purpose of the study. An n-level decomposition can be represented by a wavelet decomposition tree as in Figure 3.2A. In this dissertation, the details component obtained from the first level of decomposition was considered to be dominated by noise and was not decomposed further. The decomposition employed here can be depicted by the modified wavelet decomposition tree in Figure 3.2B.

After each level of decomposition, the length of the coefficients is reduced by a factor of two as mentioned previously. This will cause loss of time resolution. The term, time resolution, is used here in the context of wavelet theory but it is essentially spectral resolution in the context of the input signal used in this work. In our application, the use of the DWT is based on the following assumptions. First, the broad and fine (narrow) spectral features can be separated into the approximation and details coefficients, respectively. Second, all information in the original signal is represented by the collective set of wavelet coefficients. Third, the signal can be reconstructed to the same time (spectral) resolution as in the original signal by reversing the decomposition.

It is possible to reconstruct the approximations and details themselves to the original resolution by use of the obtained wavelet coefficients. For instance, reconstruction using approximation A_i yields the signal a_i by filling the details coefficients with zeros. Similarly, details D_i can reconstruct the signal, d_i. If all levels of the coefficient vectors, including $A_i, D_1, D_2, ..., D_i$ are used in the reconstruction, the original signal is obtained.

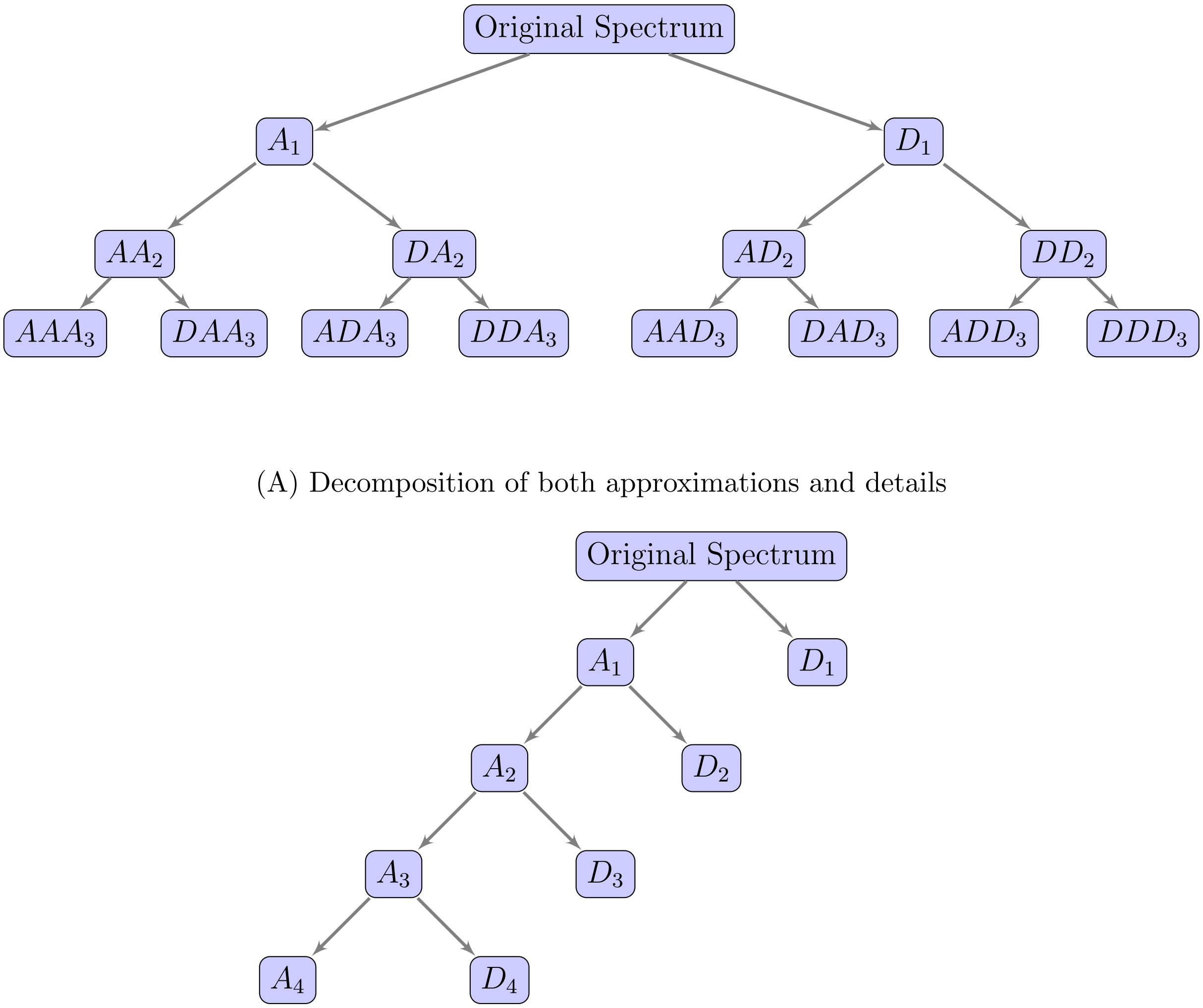
As implemented in the research described here, an input NIR spectrum is decomposed into multiple levels as shown in Figure 3.2B and then reconstructed selectively. By selecting or eliminating different levels of approximation or details coefficients, it is possible to obtain the reconstructed signal with analyte information enhanced and unwanted noise or background information eliminated. Here, the last approximation obtained in Figure 3.2B is assumed to be dominated by background information and is not included in the reconstruction. Instead, numerical optimization is used to identify which of the details to include in the reconstruction in order to provide the best suppression of unneeded spectral information. The reconstructed spectrum can then be used for quantitative analysis.

This procedure is analogous to digital filtering implemented through use of the FT (termed Fourier filtering). In Fourier filtering, the input signal is first converted into the frequency (Fourier) domain by application of the FT. Some of the frequencies can then be attenuated by multiplying the Fourier domain spectrum by the desired frequency response function of the filter. If the resulting signal is returned to the original domain by application of the inverse FT, a filtered spectrum results in which some of the underlying components have been removed. The filtered spectrum can then be used for further data analysis.

For completeness, it should be noted that the Fourier filtering procedure described above is analogous to the digital filtering methodology described previously in section 3.1.1. The multiplication of the Fourier domain spectrum by the frequency response function performed in Fourier filtering is equivalent to the convolution of the impulse response of the filter and the original NIR spectrum. This equivalence is termed the convolution theorem of the FT.

There are several families of wavelets that are especially useful in signal processing, For example, the Haar, Daubechies, Symlets and Meyer wavelet functions have been widely used.[55,57–59] Figure 3.3 depicts the shapes of the mother wavelets that define the wavelet functions employed in this work. Wavelet shapes within one family are different if in different orders as illustrated previously in Figure 3.1.

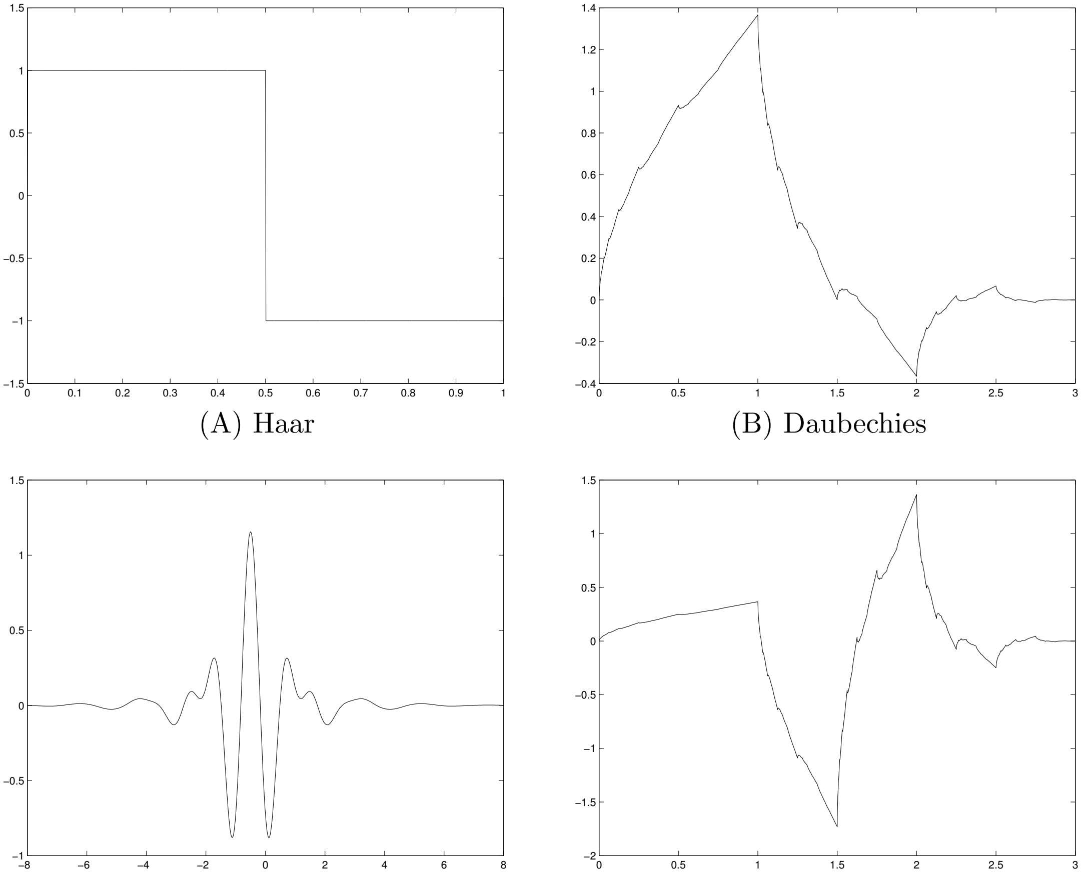
In summary, the wavelet decomposition/reconstruction process was employed here as an alternative to digital filtering for use in suppressing unwanted spectral information. The key attribute of wavelet analysis is the greater flexibility in the selection of the basis functions used in performing the spectral decomposition. Whereas digital filtering is entirely based on the use of sine and cosine functions, wavelet analysis provides a much greater selection of functions. In the dissertation research, it was hypothesized that this greater flexibility in controlling the decomposition would provide a more refined capability to extract analyte information from the measured NIR spectra.



(A) Decomposition of both approximations and details

(B) Decomposition of obtained approximations

Figure 3.2. Two types of decomposition trees of signals in wavelet analysis.



(A) Haar

(B) Daubechies

(C) Meyer

(D) Symlet

Figure 3.3. Representation of wavelet functions from different families. The caption in the subfigure shows the wavelet family.

3.2 Calibration Applications

The ultimate goal of analytical chemistry is to use the recorded signal to obtain qualitative and quantitative information regarding an analyte of interest. Examples of such information are determining the presence or absence of a particular chemical species or the amount of a species present in a sample.

To obtain knowledge regarding the concentration of an analyte, a calibration model is typically required to generate the mathematical relationship that correlates the measured signals (e.g., NIR spectra) with the parameter of interest (e.g., concentration). Calibration models can be classified as univariate or multivariate based on the dimensionality of the data input to the model. The univariate model is the simplest form, which involves relating a single measurement (e.g., the absorbance intensity at a single wavelength) with the target property of interest such as concentration. Two or more measurements are required to build a multivariate calibration model.

Multivariate calibration is widely used in current applications because of the increased power of laboratory computers, and the capabilities for rapid and large data acquisitions. Multivariate models provide increased capabilities to perform successful calibrations in complex chemical systems where there is no single measurement channel (e.g., wavelength) that provides a selective signal for the analyte. In this research, the spectral data used in calibration and prediction are all composed of multiple measurements across a given spectral bandwidth. The specific modeling techniques used in the dissertation research are described below.

3.2.1 Classical Least-Squares

Classical least-squares (CLS) analysis[3,60,61] is a multivariate calibration method based on the use of multiple linear regression (MLR) to establish the relationship between measured responses and analyte concentration. In an absorption spectroscopy application, the basic approach is to fit a measured absorbance spectrum to a model that assumes the linear additivity of a series of underlying spectral components. Here, the absorbance values in the measured spectrum across a given spectral bandwidth define the dependent variable for the fit, while the values in the known component spectra across the same bandwidth comprise the independent variables. The model is set up such that the regression coefficients obtained from the fit express the amount of each of the independent variables that are required to add together to produce the measured spectrum. One can envision that if the pure-component spectrum of the analyte is included as one of the independent variables, the regression coefficient for that term in the model will be related to the analyte concentration.

According to the Beer-Lambert law, the concentration (c) is directly proportional to the absorbance (A) as shown in Eq. 3.10, where ε is the molar absorptivity and b represents the pathlength. Absorbance and absorptivity are functions of wavelength. Assuming the absorbance of multiple components at a given wavelength is linearly additive, the Beer-Lambert law relation is shown in Eq. 3.11, in which $\mathbf{A_i}$ denotes the total absorbance at wavelength i, ε_{ij} is the absorptivity of component j at wavelength i, c_j is the concentration of the j^{th} component (l components in total in the mixture), and k_{ij} is defined as the product of ε_{ij} and b.

$$\mathbf{A} = \varepsilon bc \tag{3.10}$$

$$\mathbf{A_i} = \sum_{j=1}^{l} \varepsilon_{ij} b c_j = \sum_{j=1}^{l} k_{ij} c_j \tag{3.11}$$

For a multi-wavelength and multi-component case, Eq. 3.11 can be written in a matrix form as

$$\mathbf{A} = \mathbf{KC} + \mathbf{E} \tag{3.12}$$

where **A** is a $p \times n$ matrix constructed with the measured spectroscopic absorbance spectra of the n samples at p wavelengths each and **C** is the $h \times n$ concentration matrix with h components in the n samples. If the wavelengths are continuous and evenly spaced, **K** is a $p \times h$ matrix with the pure-component spectra of the sample constituents multiplied by the pathlength in each column. This matrix is sometimes called the sensitivity matrix. The matrix of residual spectra, **E** ($p \times n$), contains the part of the measured absorbance spectra in **A** that cannot be explained by the model. Thus, **E** contains either random noise or the spectral features of unmodeled components (i.e., spectra that are not included in **K**).

With the knowledge of the components and their pure-component spectra, the concentrations of the components can be estimated by performing MLR analysis as described above. The least-squares solution is

$$\hat{\mathbf{C}} = (\mathbf{K}'\mathbf{K})^{\mathbf{-1}}\mathbf{K}'\mathbf{A} \tag{3.13}$$

The prime symbols in Eq. 3.13 indicate the transpose of the matrix and the hat symbol on the matrix, **C**, denotes that the values are estimated rather than actual. Estimating the concentrations through this procedure is called CLS calibration.

For the case when the pure-component spectrum of each component of the sample is not available, the sensitivity matrix, **K**, in Eq. 3.12 can be estimated from the spectra of a set of mixture samples with known compositions. The number of spectra (n) used for this calculation needs to be three to five times larger than the total number of components (h) in estimating the **K** matrix. If the matrix ($\mathbf{CC}'$) is not singular, the estimation of **K** in Eq. 3.12 by least-squares is:

$$\hat{\mathbf{K}} = \mathbf{AC}'(\mathbf{C}'\mathbf{C})^{\mathbf{-1}} \tag{3.14}$$

When predicting the concentrations of unknown samples, the **K** term in Eq. 3.13 is replaced by $\hat{\mathbf{K}}$ in Eq. 3.14. Normally the error of $\hat{\mathbf{K}}$ cannot be considered negligible because it is estimated from **A** and **C**. The estimated concentrations in this case are:

$$\hat{\mathbf{C}} = (\hat{\mathbf{K}}'\hat{\mathbf{K}})^{-1}\hat{\mathbf{K}}'\mathbf{A} \tag{3.15}$$

In evaluating the quality of the calibration, the root-mean-squared error of calibration (RMSEC) error or standard error of calibration (SEC) for species j is demonstrated as

$$SEC_j = \sqrt{\frac{\sum_{i=1}^{n}(c_{j,i} - \hat{c}_{j,i})^2}{n-h}} \tag{3.16}$$

This equation assumes that n spectra have been used to estimate $\mathbf{K}$, followed by use of $\hat{\mathbf{K}}$ to estimate $\mathbf{C}$ from the same samples. The $n - h$ term in the denominator of Eq. 3.16 reflects the loss of degrees of freedom in the error estimate caused by using the same spectra to estimate $\mathbf{K}$ that are in turn used in the estimation of SEC.

When the model is used to predict the concentration of m known samples that were not used in estimating $\mathbf{K}$, the concentration error is computed as the standard error of prediction (SEP) or root-mean-squared error of prediction (RMSEP). The SEP can be obtained by

$$SEP_j = \sqrt{\frac{\sum_{i=1}^{m}(c_{j,i} - \hat{c}_{j,i})^2}{m}} \tag{3.17}$$

Here, there is no loss of degrees of freedom in determining the error estimate since the m spectra used to determine the SEP were not used in estimating $\mathbf{K}$.

The crucial factor of using CLS successfully is to implement a calibration in which the $\mathbf{K}$ or $\hat{\mathbf{K}}$ can be formulated accurately with respect to the composition of the samples to which the model will be applied. However, for complicated samples, (e.g., in environmental or biological systems), it is difficult, if not impossible, to establish all of the contributors to the overall response. In this case, the CLS method is limited in application either because of insufficient knowledge of all components or errors in the concentration estimates used in generating $\hat{\mathbf{K}}$. The latter case is especially significant when instrumental measurements are used to obtain the reference concentrations. If the precision of the spectral intensities is better than that of the concentration measurements, the formulation of the least-squares model in Eq. 3.15 is technically invalid because the independent variables now have greater error than the dependent variable.

To address the limitations noted above, Haaland and coworkers have developed a series of strategies to add flexibility to the CLS model.[62–65] Termed augmented CLS (ACLS), the concept is to expand $\mathbf{K}$ in Eq. 3.13 to include additional "spectral shapes" that do not technically correspond to pure-component spectra of chemical constituents of the sample. For example, if baseline variation typically exists in the measured spectra due to instrumental drift effects, the shapes of baseline components can be added as additional spectra in $\mathbf{K}$ to form a new augmented $\mathbf{K}$, $\mathbf{K_a}$. When the least-squares fit in Eq. 3.13 is performed, the contribution of the baseline component to the measured absorbance spectrum is then taken into account, resulting in more accurate concentration estimates for the actual chemical components. Stated differently, the least-squares model now no longer has to adjust the regression coefficients (concentration estimates) of the chemical components to account for the baseline contribution to the measured spectrum. Other approaches can be used to augment $\mathbf{K}$, including composite spectra to describe components of the sample matrix that do not change or change together from sample to sample. The ACLS method is used in Chapters 6 and 7. The specific implementation used in the dissertation research will be discussed there.

3.3 Multiple Linear Regression Models

To handle the case in which the full matrix of sample components cannot be specified accurately, an inverse regression model can be employed in which the concentration becomes the dependent variable ("y") and the measured spectral intensities define the independent variables ("x"). Then, with the knowledge of the measured spectral intensities and concentrations of a set of calibration samples, the concentration is modeled as a function of the measured instrumental responses by MLR. This is also called an inverse calibration method.[3,61,66,67]

In a multivariate inverse calibration model, for a component in a sample, the concentration of the analyte of interest can be modeled as a function of the spectroscopic intensities at multiple wavelengths as:

$$c_i = b_0 + b_1 x_{1,i} + b_2 x_{2,i} + ... + b_p x_{p,i} + e_i \tag{3.18}$$

where c_i is the concentration of the analyte of interest in sample i, $x_{1,i}, x_{2,i}$, ..., $x_{p,i}$ represent the spectral intensities from p wavelengths, and b_0, b_1, b_2, ..., b_p, are the regression coefficients returned by MLR. The p wavelengths could be selected from a total of m depending on a wavelength selection procedure. For a series of measured concentrations of h components and their corresponding instrumental responses, Eq. 3.18 can be written in matrix form as

$$\mathbf{C} = \mathbf{X}\mathbf{B} \tag{3.19}$$

where **C** is an $n \times h$ matrix which records the concentrations of h components from n samples, **X** represents the $n \times p$ matrix with the spectral intensities from p wavelength, and **B** $(p \times h)$ is the regression coefficient matrix. The **B** matrix can be estimated by the generalized inverse:

$$\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{C} \tag{3.20}$$

Given **B**, predicted concentrations in new samples with spectra in **A** can be estimated by

$$\hat{\mathbf{C}} = \mathbf{A}\mathbf{B} \tag{3.21}$$

Standard errors in the predicted concentrations are then estimated by Eq. 3.16 for the calibration samples and Eq. 3.17 for samples not included in the calibration matrix of data used to compute **B**. The denominator in Eq. 3.16 is changed to $(n - p - 1)$ in this case to reflect the correct number of degrees of freedom ($p + 1$ model terms in Eq. 3.18).

As in the case discussed previously for estimating **K** in the CLS method, n must be sufficiently larger than p to allow the precise determination of **B**. The American Society of Testing and Materials (ASTM) standard is $(n > 6p)$.[68] For example, if the spectral scan is from 4800 to 4200 cm^{-1} with a 4 cm^{-1} point spacing, there are 151 responses in each spectrum. By the ASTM standard, $6 \times 151 = 906$ samples are thus required to build the calibration model in this case. This number of samples is not practical in most applications. Therefore, wavelength selection methods must be applied to identify the key spectral points for use in building the calibration model. Unfortunately, because of the large amount of spectral overlap that often occurs in mixture samples, it can be difficult to identify a small subset of wavelengths that will build an adequate calibration model.

Another consideration in this calibration method is the condition of $\mathbf{X}'\mathbf{X}$ in Eq. 3.20. The inverse of $\mathbf{X}'\mathbf{X}$ is required to calculate the regression coefficients. If the columns or rows are linearly dependent, the $\mathbf{X}$ matrix is singular or nearly so and thus poorly conditioned for the inverse calculation. Unfortunately, such collinear relationships are common with spectroscopic data where multiple spectral points typically are acquired across each spectral band. This factor, coupled with the difficulty of identifying small subsets of wavelengths that will adequately model the concentrations, has led this calibration method to be replaced with a modified approach described in the next section.

3.4 Latent Variable Methods

The idea of latent variable methods is to compute a new set of independent variables (orthogonal to each other) which are linear combinations of the original responses. An inverse regression model of the type described in Eq. 3.18 is then built to relate concentrations to the new responses obtained from the latent variables. The goal of this approach is to extract the key spectral information in fewer points than the raw data. This allows the calibration model to be built with fewer terms and thus satisfy the ASTM standard with fewer calibration samples. Furthermore, since the latent variables are orthogonal, there are no issues with the inversion of $\mathbf{X}'\mathbf{X}$.[3,61]

After obataining the latent variables, the following relationship can be obtained:

$$\mathbf{T} = \mathbf{RS} \tag{3.22}$$

In Eq. 3.22, $\mathbf{R}$ is the $n \times p$ matrix of original measured data, $\mathbf{S}$ $(p \times h)$ represents the h latent variables computed from the data, and $\mathbf{T}$ is an $n \times h$ matrix of scores. The columns of $\mathbf{S}$ have the same dimensionality as the original spectra and are called factors or loadings. The loadings can be considered basis vectors in an h-dimensional coordinate system, while the scores represent the projections (coordinates) of the original spectra onto this new basis.

With the h-dimensional score vector corresponding to each input spectrum, the calibration model can be built in a manner analogous to Eq. 3.18. The computed scores are used as the new independent variables.

$$c_i = b_0 + b_1 t_1 + b_2 t_2 + ... + b_h t_h \tag{3.23}$$

The key to the success of latent variable methods is the calculation of the loadings such that the corresponding scores efficiently represent the key spectral information needed to build an effective calibration model. The two most common approaches to this calculation are described below.

3.4.1 Principal Component Regression

Principal component regression (PCR) is simply principal component analysis (PCA) followed by an MLR step.[3,61] The goal of PCA is to factorize the response matrix into h factors as shown below:

$$\mathbf{R} = \mathbf{T}\mathbf{V}' + \mathbf{E} \tag{3.24}$$

In Eq. 3.24, $\mathbf{R}$ is the response matrix with n spectra in the rows and p wavelengths, $\mathbf{T}$ is the score matrix with dimensionality of $n \times h$, $\mathbf{V}$ is a $p \times h$ matrix with h orthonormal loadings in the columns, and $\mathbf{E}$ is the matrix of residual spectra that contains the unmodeled portion of $\mathbf{R}$. The loading matrix, $\mathbf{V}$, contains h of the p eigenvectors of $\mathbf{R}'\mathbf{R}$. The eigenvectors can be obtained through singular value decomposition (SVD) or the nonlinear iterative partial least-squares (NIPALS) algorithm.[69,70]

Each loading, called a principal component (PC), has an associated eigenvalue that is proportional to the magnitude of the loading vector before normalization. The loading with the largest eigenvalue projects most strongly onto $\mathbf{R}$ and is called the first principal component. Because the loadings are orthogonal, each projects onto (i.e., accounts for) a unique portion of the information in $\mathbf{R}$. By taking the loadings corresponding to the h largest eigenvalues, an h-dimensional basis is obtained that can most efficiently represent the information in **R**.

The score matrix is computed from **R** and **V**. For eigenvector i, the corresponding score vector is

$$\mathbf{t}_i = \mathbf{R}\mathbf{v}_i \tag{3.25}$$

Since **V** is an orthogonal matrix, the calculated score vectors are also orthogonal. Thus, assuming there is significant collinearity present in **R**, the scores offer an ability to represent the key information in **R** while requiring fewer variables to do so. This leads to a calibration model with fewer terms (Eq. 3.23).

In building the calibration model in PCR, Eq. 3.19 is employed by replacing **X** with the score matrix, **T**. For a single analyte, the regression coefficient vector, **b**, can be computed by substituting **X** with **T** in Eq. 3.20, where **c** in Eq. 3.26 is the vector of reference concentrations for the calibration samples.

$$\mathbf{b} = (\mathbf{T}'\mathbf{T})^{\mathbf{-1}}\mathbf{T}\mathbf{c} \tag{3.26}$$

Once the regression coefficients have been computed by use of Eq. 3.26, a calibration model is obtained that can be used with future samples. For a concentration to be predicted, the recorded spectrum is used together with the **V** matrix computed from the calibration data. This allows Eq. 3.25 to be used to compute the corresponding score vector. The predicted concentrations are then computed by Eq. 3.23.

The PCR method is based on the assumption that the major sources of variance in **R** arise from the instrumental signals and that the measurement of concentration in the calibration samples is accurate and precise. If these assumptions are not valid, the model will likely require a large number of terms and the obtained regression coefficients will not be reliable for use in predicting concentrations of future samples.

The key challenges in the practical use of PCR are to define the spectral region that is to be used in the calculation of the loading vectors and to determine h, the optimal number of loadings to use in the calibration model. If spectral wavelengths are included in the PCA calculation that are high in noise or that contain information irrelevant to the analyte determination, the computed loadings will be sub-optimal for use in modeling the analyte concentrations. This problem arises from the fact that the loading vectors that explain the largest sources of variation in the input data are selected for use in the calibration model. If the spectral variance arising from changes in analyte concentration represents a minor component of the total variance, PCA may not necessarily extract the specific analyte information efficiently.

In practical use with spectroscopic data, h will typically be much less than p, the number of spectral points in the response data submitted to PCA. Determining the best value of h is an optimization issue, as there is no theoretically best value. Various strategies can be employed to determine the optimal value of h.[71] In the dissertation research, selection of h in the latent variable models was incorporated into the model optimization step. The specific approach used with each data set will be described in the relevant results chapters.

3.4.2 Partial Least-Squares Regression

A limitation of PCA is that decomposition of the input data is based solely on the explanation of variance. The selected principal components will be those that explain the most spectral variance, not necessarily those that explain the information most useful in modeling changes in analyte concentration. The partial least-squares (PLS) regression method is a related latent variable technique that attempts to address this limitation. The PLS method has been widely applied in analytical chemistry.[3,61,72]

Compared to PCA, instead of modeling the experimental response exclusively, PLS also takes the concentration of the analyte of interest into account in generating the orthogonal latent variables. For the calibration data, the obtained components maximize the covariance between the spectra and the reference concentrations. The joint decomposition of the spectral matrix, $\mathbf{X}$, and the concentration vector, $\mathbf{c}$, can be implemented by

$$\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E} \tag{3.27}$$

$$\mathbf{c} = \mathbf{T}\mathbf{q} + \mathbf{e} \tag{3.28}$$

The matrix, $\mathbf{X}$, in Eq. 3.27 contains the n measured calibration spectra in the rows (p wavelengths), while $\mathbf{c}$ in Eq. 3.28 is the $n \times 1$ concentration vector of the analyte of interest. The matrices, $\mathbf{T}$, $\mathbf{P}$, and $\mathbf{E}$, are analogous to $\mathbf{T}$, $\mathbf{V}$, and $\mathbf{E}$ in Eq. 3.24.

In terms of dimensionality, $\mathbf{T}$ is the $n \times h$ score matrix and $\mathbf{P}$ is the $p \times h$ loading matrix, whose rows are the empirically derived latent variables (also called spectral loadings).

In PLS, the concentration vector $\mathbf{c}$ is also decomposed into matrix $\mathbf{T}$ and vector $\mathbf{q}$ as shown in Eq. 3.28. The $\mathbf{T}$ matrix is the same as the one in Eq. 3.27. The vector $\mathbf{q}$ $(h \times 1))$ is analogous to a loading vector of the concentrations. After $\mathbf{X}$ and $\mathbf{c}$ are decomposed into the h latent variables, the remaining information is found in the spectral residual matrix, $\mathbf{E}$, and the concentration residual vector $\mathbf{e}$. The scores are orthogonal, but the spectral loadings ($\mathbf{P}$) are not orthogonal and they are normally not normalized. In PLS, the scores and loadings are dependent on both the instrumental responses and concentrations of analyte, which is different from PCA.

There are various algorithms to link $\mathbf{X}$ and $\mathbf{c}$ in calculating the scores and loadings.[72,73] In this thesis, the decomposition of $\mathbf{X}$ and $\mathbf{c}$ is obtained by computing a set of loading weights, $\mathbf{w}$, for each spectral loading. First of all, $\mathbf{X}$ and $\mathbf{c}$ are mean-centered. The first loading weight vector $\mathbf{w}_1$ is calculated as

$$\mathbf{w}_1 = \frac{\mathbf{X}'\mathbf{c}}{\|\mathbf{X}'\mathbf{c}\|} \tag{3.29}$$

As shown in Eq. 3.29, $\mathbf{w}_1$ is normalized to the unit length by $\|\mathbf{X}'\mathbf{c}\|$. With $\mathbf{w}_1$ as a basis vector, the first score vector ($\mathbf{t}_1$) is defined as the projection of $\mathbf{X}$ on $\mathbf{w}_1$ as shown in Eq. 3.30, and its corresponding spectral loading ($\mathbf{p}_1$) is computed as in Eq. 3.31. Similarly, the first concentration loading is given by Eq. 3.32. These calculations are least-squares fits of the spectral matrix and concentration vector onto the computed scores.

$$\mathbf{t} = \mathbf{X}\mathbf{w}_1 \tag{3.30}$$

$$\mathbf{p}_1 = \frac{\mathbf{X}\mathbf{t}_1}{\|\mathbf{t}_1'\mathbf{t}_1\|} \tag{3.31}$$

$$\mathbf{q}_1 = \frac{\mathbf{t}_1'\mathbf{c}_1}{\|\mathbf{t}_1'\mathbf{t}_1\|} \tag{3.32}$$

After the scores and loadings are computed for the first latent variable, the residuals of the spectral matrix and concentration vector are calculated, respectively.

$$\mathbf{E}_1 = \mathbf{X} - \mathbf{t}_1\mathbf{p}_1' \tag{3.33}$$

$$\mathbf{e}_1 = \mathbf{c} - \mathbf{t}_1\mathbf{q}_1 \tag{3.34}$$

The residuals express the remaining spectral information ($\mathbf{E_1}$) and concentration information ($\mathbf{e_1}$) which was not extracted into $\mathbf{t_1}$.

Next, the second vector of loading weights, scores, and loadings are calculated by Eqs. 3.30 - 3.32 by simply replacing $\mathbf{X}$ and $\mathbf{c}$ with $\mathbf{E_1}$ and $\mathbf{e_1}$. After that, the new spectral and concentration residuals are computed by Eqs. 3.33 and 3.34. This procedure is repeated until the scores and loadings of the h latent variables are obtained.

In PLS, using the loading weight vectors, $\mathbf{w}$, is important in calculating the scores, $\mathbf{t}$, because it incorporates the information from the analyte concentration into the calculation of the PLS factors. Therefore, PLS is a bilinear latent variable model. The extraction of information from $\mathbf{X}$ is biased to explain the information in $\mathbf{c}$, rather than solely to explain the variance in $\mathbf{X}$.

Once the scores are obtained, the procedure of building the calibration model with the PLS method is identical to that presented previously for PCR. For the prediction of unknown concentrations, the response matrix of unknown spectra ($\mathbf{R}_{pred}$) is first centered with the mean of the calibration data and the prediction score matrix is computed with the first vector of loading weights computed previously from the calibration data. Then, the contribution of $\mathbf{t}_{1,pred}$ is removed from the response matrix by use of the previously computed first spectral loading. These calculations are summarized in Eqs. 3.35 and 3.36

$$\mathbf{t}_{1,pred} = \mathbf{R}_{pred}\mathbf{w}_1 \tag{3.35}$$

$$\mathbf{R}_{pred,1} = \mathbf{R}_{pred} - \mathbf{t}_{1,pred}\mathbf{p}_1' \tag{3.36}$$

This process continues to obtain all the required scores. The unknown concentration can then be calculated by use of the previously computed regression coefficients. A mathematically equivalent way of prediction is through the use of a regression coefficient vector as specified in Eq. 3.37.

$$\mathbf{b} = \mathbf{W}(\mathbf{P}'\mathbf{W})^{-1}\mathbf{q} \tag{3.37}$$

$$\hat{\mathbf{c}} = \mathbf{R}_{pred}\mathbf{b} + \bar{\mathbf{c}} \tag{3.38}$$

In the equations above, $\mathbf{W}$ is a $p \times h$ matrix with the columns containing the loading weights and $\mathbf{P}$ is a $p \times h$ matrix in which the columns are the spectral loadings. The elements of $\mathbf{q}$ are the concentration loadings, and $\bar{\mathbf{c}}$ is the mean concentration from the calibration data.

The same issues discussed previously with respect to optimization of the number of latent variables and the spectral range pertain to the PLS method. These decisions are often more difficult with PLS than with PCA. Because the concentrations are used in the extraction of the spectral loadings, there is the likelihood that random noise or spectral artifacts that happen to correlate with the concentration vector will be extracted from the calibration data matrix. These factors may appear to be significant when building the regression model, but they will not be useful when the model is applied to future data. Both the spectral range submitted to the PLS calculation and the number of latent variables used in the calibration model must be studied carefully during the model optimization process.

3.5 Model Validation

The ultimate goal of building the calibration model is to predict the analyte concentrations for data collected in the future. Hence, model validation is an essential step in model development to help ensure good prediction performance. Validation of the calibration model can be applied with either internal or external approaches.

3.5.1 Internal Validation

With internal validation, a model is tested within the calibration data. The value of SEC (Eq. 3.16) can provide a rough estimation of model performance. However, especially in PCR and PLS approaches, calibration performance can always be improved by increasing the model size; therefore, it would be misleading if the model were optimized based on the calibration performance alone. Two approaches to internal validation will be discussed: (1) use of a monitoring set and (2) cross-validation.

A subset of the calibration data can be selected to evaluate the prediction performance of the model. This data set is called the "monitoring data". The monitoring data set is withheld from the calculation of the calibration model and is used subsequently to test the computed model. This approach assumes the data in the monitoring set will resemble the future data to which the model will be applied.

One way to pick the monitoring set is to split the calibration set into subsets of calibration and monitoring. The splitting can be done randomly or on the basis of the sequence of the data collection. As noted above, the calibration model is constructed first with the calibration subset, and the model is then applied to predict the concentrations corresponding to the spectra in the monitoring set for the purpose of evaluating prediction performance. The SEP calculation from Eq. 3.17 can be applied to the monitoring data to produce the standard error of monitoring (SEM).

The SEM value estimates the prediction performance that would be obtained for data that are consistent with the spectra in the calibration set. Consequently, it can serve as a criterion for use in optimizing the parameters of the calibration model such as the number of latent variables or the wavelengths submitted to PCA or PLS.

Especially for the PLS method, adding new factors to the model will always decrease the calibration error. However, unnecessary factors could introduce irrelevant information into the model and may cause a poor prediction for data outside of the calibration set. Stated differently, selection of too many factors in the calibration model may make the model lose its ability to generalize to samples collected in the future. A traditional procedure of the optimization is to systematically increase the number of latent variables and calculate the SEM values for each level. Typically, the SEM values will decrease as the model size increases, and taper off (perhaps increase) as the model size becomes greater than required. The rule of model size optimization is to select the fewest number of factors that provide an adequate SEM value. An F-test is often employed to choose the model size that provides a value of SEM that is not statistically different from the minimum SEM.

Instead of selecting one subset of the calibration data for use in evaluating prediction performance, cross-validation cycles the calibration set through calibration/monitoring subsets such that each calibration spectrum is withheld from the model and predicted once. The size of the monitoring subset can vary from a single spectrum (leave-one-out cross-validation) to any larger fraction of the calibration set. When subsets of size greater than one are used, a decision has to be made as to how to sample the subsets. The subsets can be contiguous blocks of the calibration set, randomly picked subsets, or subsets defined through a structured selection pattern (e.g., every 10th spectrum).

The predicted concentrations obtained from cycling through the subsets are pooled and an overall prediction error is computed. The calculation is the same as in Eq. 3.17, with the result typically called the cross-validated SEP (CV-SEP). The CV-SEP value is then used as described above for the selection of the number of model terms or for the optimization of other model parameters.

3.5.2 External Validation

With internal validation, using the monitoring subset selected from the calibration set may result in an overly optimistic evaluation. In predicting data collected in the future, however, if time-correlated issues, such as instrumental drift or environmental changes (e.g., temperature drift) are present, the calibration model may perform poorly since additional factors may be missed. Therefore, an external validation is often necessary to help evaluate such issues and obtain a realistic estimate of model performance. With this approach, an external data set collected outside of the time frame of the calibration data is used to evaluate the calibration model performance. The SEP is used to evaluate performance instead of the SEM.

3.6 Optimization Methods

As discussed previously, between the choice of preprocessing parameters and parameters associated with the calibration model calculation, there are many optimization decisions that must be made in order to obtain a well-performing calibration model. Optimization can be performed by assigning discrete levels to each of the variables to be optimized and then evaluating each combination of parameter values through calculation of SEM, CV-SEP, or SEP. This method is termed a grid search. It is a straightforward procedure but is limited in that the use of a large search space (i.e., many variables and many levels) is computationally impractical.

In this research, a numerical optimization procedure called particle swarm optimization (PSO) was employed. Numerical optimization methods allow the definition of a large search space and then use various algorithmic strategies to navigate the search space in search of the optimal combination of parameters. This is done without evaluating every combination of parameter values and is thus computationally more attractive if a large search space needs to be interrogated. The PSO method is outlined below.

3.6.1 Particle Swarm Optimization

The PSO method is a population-based stochastic optimization method developed by Eberhart and Kennedy in 1995.[74–76] This optimization technique was discovered through simulation of a social behavior model, e.g., fish schooling, bird flocking, or people working together. It is based on the assumption of social information exchange. For example, to solve a problem as a team, people would exchange information, beliefs and attitudes with each other. The interaction could help them in moving towards the correct solution, and hence enable them to solve the problem faster. The particles in PSO are metaphors of a group of people. They fly in the search space, adjust locations and directions and try to find the optimal solution by exchanging information.

The PSO technique has many similarities with the genetic algorithm (GA)[77,78] and other evolutionary computational methods. Such evolutionary optimization methods start with an initial population with random solutions, determine how good the solutions are by use of a fitness function, and search for the optima by updating the solutions over generations. The PSO algorithm has been successfully applied in various application areas for optimization of a wide range of continuous function.[75,79,80] Next, a brief introduction of the PSO implementation used in the dissertation research is provided.

The PSO implementation divides into three steps, initialization, evaluation and update. In the initialization step, PSO creates an initial population of particles with random positions $\vec{x}_i$ and velocities $\vec{v}_i$. Each particle is a potential solution of the problem. The dimensionality of the particles is determined by how many parameters are optimized. The positions and the velocities of the particles are represented numerically as vectors of integers, which are the indices of the mapped possible values of the parameters.

The second step involves the evaluation of the position of each particle with a fitness function. The fitness value is computed by the fitness function to judge how good the position of the particle is as a solution. The fitness value is the criterion used to guide the particles moving toward the global optimum. For each particle, the best solution achieved so far is stored and its fitness value is called its *pbest*. Another ‘best’ value tracked by PSO is the best value obtained so far by the neighbors of the particle. It is called *lbest*. When a particle takes all of the population as its topological neighbors, the best value becomes the global best and is called *gbest*. The *pbest* and *gbest* and their corresponding locations are recorded for each generation.

At the third step, all particles are updated to a new position and velocity. When flying through the search space, each particle keeps track of its coordinates in the space. The basic concept of the updating strategy is that each particle is accelerating toward its *pbest* and the *gbest* locations with a random weighted acceleration. Both the velocities and positions are updated based on the *gbest* and *pbest* of each particle. At each generation, the next position of each particle is calculated by the current position and the new velocity. The new velocity value of each particle is determined based upon the current velocity, the distance between the current position and *pbest* and the distance to *gbest*. Eqs. 3.39 and 3.40 show the basic algorithm.

$$\vec{x}_i^{k+1} = \vec{x}_i^k + \vec{v}_i^{k+1} \tag{3.39}$$

$$\vec{v}_i^{k+1} = a \times \vec{v}_i^k + b_1 \times r_1 \times (\vec{p} - \vec{x}_i^k) + b_2 \times r_2 \times (\vec{g} - \vec{x}_i^k) \tag{3.40}$$

In Eqs. 3.39 and 3.40, $\vec{x}_i^{k+1}$ and $\vec{x}_i^k$ represent the positions of particle i at iterations $k+1$ and k, respectively. Analogously, $\vec{v}_i^k$ and $\vec{v}_i^{k+1}$ are the corresponding velocities. The term $\vec{p}$ is the *pbest* of particle i while $\vec{g}$ represents the *gbest*, which is the global best location in the whole swarm of particles. In updating the velocity, the coefficient a is an inertial constant, which specifies how much the velocity in the next generation is affected by the current velocity. Coefficients b_1 and b_2 are constants to define the strength of attractions of *pbest* and the *gbest* to the particle. The random values, r_1 and r_2, are used to introduce a degree of randomness in the process. They are randomly selected between [0, 1] from the uniform distribution. Introducing a random component into the optimization can help to avoid local optima when exploring the search space. The two equations indicate that each particle adjusts its position based on the two positions (*pbest* and *gbest*) when moving through the search space. The acceleration process of a two-dimensional particle is shown in Figure 3.4.

The evaluation and updating steps are then repeated for a specified number of generations or until the best location (best fit) has been found. The PSO method was used as part of the calibration development work in Chapters 5 - 7. Specific details regarding the variables optimized and the fitness function used will be provided in the relevant sections of those chapters.

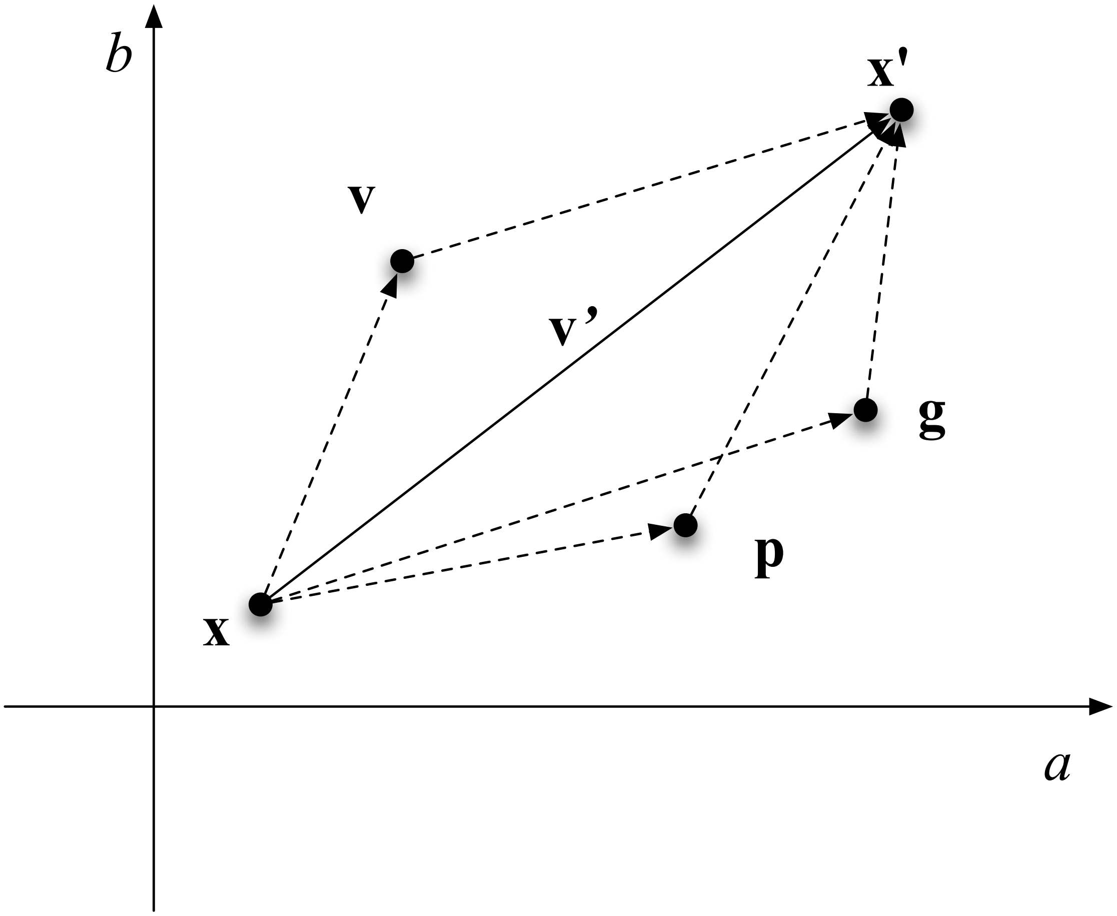


Figure 3.4. An example of the acceleration process of the particles in the updating step of PSO in a two-dimensional search space based on variables a and b to be optimized. The values, **x** and **x'**, are the initial and updated location vectors of one particle, respectively, while **v** and **v'** are the corresponding velocity vectors. Point **p** is the current personal best location and point **g** is the global best location.

CHAPTER 4
QUANTITATIVE DETERMINATION OF METHANOL AND ETHANOL WITH SYNTHETIC CALIBRATION SPECTRA IN PASSIVE FOURIER TRANSFORM INFRARED REMOTE SENSING MEASUREMENTS

4.1 Introduction

Fourier transform infrared (FT-IR) remote sensing is used to monitor atmospheric species between the spectrometer and an IR source. Depending on the IR source type, remote sensing can be divided into active and passive modes. Measurements which rely on an instrumentally controlled IR source are categorized as active mode, while those that view the uncontrolled (i.e., naturally occurring) background IR radiation present in the scene are termed passive measurements. The passive configuration is much more operationally versatile, because no controlled IR background is commonly encountered in either ground-based or airborne measurement scenarios. This technique has been used in various atmospheric monitoring applications because of its capability to analyze a large volume of the atmosphere without sample collections. It offers applications in stack emission analysis, detecting toxic gases in the workplace, and leak detection.[81,82]

In the passive mode experiment, a ground-based emission spectrometer configured with telescope-enhanced entrance optics allows the collection of IR radiance within the field of view (FOV) against sky, terrestrial or manmade backgrounds. The spectrometer will collect the IR radiance emitted from the background, analyte cloud and atmospheric gases in the FOV. In terms of atmospheric gases, carbon dioxide, water vapor and ozone can be interfering components. Emission or absorption features of the analyte depend on the temperature difference between the analyte (e.g., the gaseous effluent from a stack) and the background. If the background is at a lower temperature than the analyte, emission spectra will be obtained. With the opposite condition, absorption occurs. According to Planck's function, temperature is the major factor in determining the radiance intensity. Therefore, a significant temperature differential between the analyte plume and the background is the primary requirement for detection capability.[83] The greater the differential temperature is, the more analyte signal can be observed superimposed on the background signal.

Usually in remote sensing measurements, the concentration and path length are taken as a combined value, because there is no path length measurement (l) offered for gas samples during an outdoor data collection. Also, it is impossible to differentiate a narrow analyte cloud with a high concentration (c) from a wide analyte plume with a low concentration. The combined product (cl) is used to represent the abundance of the sample. The product value is termed path-averaged concentration with the typical units of part-per-million-meter (ppm-m).

In the passive detection mode, the radiance emitted from the background is constantly varying, as the temperature of the background within the FOV is uncontrolled. If the spectrometer is placed on a moving platform such as an aircraft, the background scene is also undergoing constant change, resulting in a corresponding change in emissivity. Emissivity is a parameter varying from zero to one that describes the degree to which a material behaves as a theoretical blackbody. Materials with an emissivity of one behave as a blackbody, while those with emissivities less than one are termed graybodies. Even in a static measurement application, the occurrence of moving objects in the background scene can lead to variation in emissivity. In addition, the composition of the intervening atmospheric gases between the spectrometer and background is constantly changing because of changing meteorological conditions or spectrometer movement. All of these factors combine to produce an unstable background radiance. This variance in the background radiance makes it impossible to obtain stable reference background spectra and also stable sample spectra. The traditional laboratory spectral processing approach based on taking the ratio of a sample spectrum to a stable background spectrum is not applicable.

These factors make the development of a successful quantitative analysis procedure for passive IR data extremely challenging. The principal difficulties are (1) to extract reliable quantitative information from the measurement and (2) to correlate spectral intensities to analyte concentrations through a calibration procedure. Signal processing methods (e.g., digital filtering) have been investigated to assist the information extraction step by removal of sources of data variation that are unrelated to changes in the concentration of a target analyte.[11,44,83,84] The calibration procedure that relates signal intensity to concentration is complicated by the high cost of conducting controlled release measurements in the outdoor environment, as well as by the significant sources of uncertainty inherent in such measurements. Outdoor releases of toxic chemicals have additional impediments to practicality.

In the work described in this chapter, a data synthesis approach is used to develop calibration data sets for use in building quantitative models based on passive IR remote sensing data. The method of using synthetic calibration spectra based on a linear radiometric model has been successfully applied in both the spectral and interferogram domains.[83–85] Using synthetic spectral data enables flexibility in changing key input parameters such as the chemical components (analytes or interferences) in the FOV, analyte concentrations, and the temperatures of the background and target cloud. When building a calibration model, similarity between the calibration data and any prediction data to which the model will be applied is important. If the conditions associated with the prediction data are known, the simulation approach is beneficial in flexibly generating calibration data that are similar to the prediction data.

In this chapter, a quantitative analysis method for pure and mixture samples of ethanol and methanol is developed by use of simulated calibration data computed through appropriate radiometric models. Gas spectra collected in the laboratory from a static gas cell and field data collected during an outdoor stack emission monitoring experiment were studied. Due to instabilities in the experimental conditions, the initial part of the experimental data needed to be used to determine certain parameters associated with a given measurement block. After that, synthetic calibration spectra were generated to build partial least-squares (PLS) models for quantification of subsequent samples.

4.2 Theory

4.2.1 Radiance Model

The work described here involved the use of a spectral synthesis procedure to obtain data for use in building quantitative models for predicting analyte concentration from passive IR spectra. This required the adoption of an appropriate radiometric model for use in simulating the laboratory and field data employed in this work. Details about the radiance model applied here have been introduced in Chapter 2.

The basic setup of the model assumes a target gas viewed against a background blackbody or graybody scene. It is assumed that the target vapor fills the spectrometer field-of-view (FOV) and that any light loss due to scattering can be ignored. The target gas contributes to the signal received by the sensor by either absorbing light emitted by the background or by acting as a selective radiator and thereby directly emitting light. In this model, the spectral radiance in the FOV at a given wavelength can be written as

$$L_x = \tau_a \tau_t L_{bkg} + (1 - \tau_a \tau_t) L_t \tag{4.1}$$

In Eq. 4.1, L_x represents the total spectral radiance emanating from the scene, L_{bkg} is the radiance of the background at a temperature of T_{bkg}, and L_t is the radiance estimated by Planck's function at the temperature of the target gas, T_t. The radiance values are modified by τ_a and τ_t, the transmittance values of the intervening atmosphere and the target vapor, respectively. In the L_{bkg} term, the transmittance values are multiplied by the background radiance to encode the absorption component of the received signal. In the L_t term, $(1-\tau_a\tau_t)$ represents the emittance of the target gas. Multiplication by L_t specifies the emission component of the signal.

The transmittance of the target gas (τ_t) is defined as $\tau = \exp(-\alpha cl)$, where α is the absorptivity of the analyte at the specified wavelength (m^2/mg), c is the concentration of the gas (mg/m^3), and l is the optical pathlength or depth of the cloud along the optical axis (m). The absorbance (A) of the target cloud can be calculated by $A = 0.434(\alpha cl)$, where cl is the path-averaged concentration noted previously.

For the work described here, the emissivity of the background will be assumed to be 1.0 in all cases, thereby allowing it to be modeled with Planck's function. In addition, the atmospheric transmittance term will be ignored (i.e., $\tau_a \approx 1$) because of the small distance between the background and spectrometer in the data sets employed in this work. Thus, for a given target gas, the radiance received by the sensor will be assumed to depend solely on the background and target gas temperatures, the concentration of the gas, and the cloud depth.

4.2.2 Generation of Synthetic FT-IR Spectra

The synthesis of passive single-beam FT-IR spectra was performed according to Eq. 4.1. Employing the assumption that the emissivity equals unity, the spectral radiance can be computed from Planck's blackbody equation:

$$L^*(\vec{\nu}, T) = \frac{C_1 \times \bar{\nu}^3}{\exp(\frac{C_2 \times \bar{\nu}^2}{T}) - 1} \tag{4.2}$$

where $C_1 = 2hc^2 = 1.191 \times 10^{-12}\mathrm{W/cm^2 sr(cm^{-1})^4}$ and $C_2 = hc/k = 1.439\mathrm{K} \cdot \mathrm{cm}$. The values, C_1 and C_2, are termed the first and second radiation constants, $\bar{\nu}$ is the wavenumber of the radiance, and T is the temperature of the blackbody. From this equation, for the spectrum of a perfect blackbody material, the only parameter that can affect the spectral intensity is the temperature.

Besides the spectral radiance received from the scene within the FOV, the instrument itself also contributes to the recorded single-beam spectrum. In computing the spectrum, specific information about the instrument must be taken into consideration in the form of the detector response and the instrument self-emission function.

As a function of wavenumber, the single-beam spectrum recorded by the spectrometer while viewing the input radiance, L_x can be written mathematically as defined by Shaffer and Combs:[83]

$$S = r(L_x + L_e) \tag{4.3}$$

In Eq. 4.3, S is the computed single-beam spectrum, r is the instrument responsivity, and L_e is the instrument self-emission function, also called the instrument offset. The responsivity or gain primarily measures the sensitivity of the detector at each wavenumber. The self-emission term arises from both emission and scattering contributions of various components in the optical train and becomes significant because the temperature of the ambient background is similar to the internal temperature of the instrument. The two instrumental terms in Eq. 4.3 are temperature-dependent and thereby result in instability during the operation of a passive FT-IR spectrometer.

Calculation of an accurate single-beam spectrum requires that r and L_e be defined. In a stationary spectrometer configuration, this is done by measuring a reference IR blackbody source at "hot" and "cold" temperatures. Then, r and L_e are calculated according to Ballard[86] as

$$r = (S_h - S_c)/(L_h^* - L_c^*) \tag{4.4}$$

$$L_e = [(S_c \times L_h^*) - (S_h \times L_c^*)]/(S_h - S_c) \tag{4.5}$$

In Eqs. 4.4 and 4.5, S_h and S_c are the single-beam spectra for the hot and cold blackbody source collected by the FT-IR instrument, and L_h^* and L_c^* are the radiances predicted by Planck's function at the same hot and cold temperatures, respectively. Assuming a linear detector response, r and L_e can be computed. Subsequently, Eqs. 4.1 and 4.3 can be used together to produce a synthetic single-beam spectrum that incorporates the desired values of $T_{bkg}, T_t,$ and cl.

4.3 Experimental

4.3.1 Instrumentation

4.3.1.1 Laboratory Data

The instrumental setup of the laboratory data collection in this study is shown in Figure 4.1A. Interferogram data were collected with a Midac FT-IR emission spectrometer (Model M2411, Midac Corp., Westfield, MA) equipped with a liquid nitrogen-cooled Hd:Cd:Te (MCT) detector. Interferograms of 1024 points were collected as single scans and sampled at every eight zero-crossings of the internal He-Ne reference laser. The maximum spectral frequency was 1974.75 cm^{-1} and the nominal spectral resolution was 8 cm^{-1}. Interferograms were collected at approximately 1.6 scans per sec.

Background radiance in this experiment was provided by a 14 × 14 inch blackbody source (Electro Optical Industries, Inc., Santa Barbara, CA). The device was thermoelectrically controlled in order to allow both cooling and heating relative to ambient temperature. This allowed simulating both emission and absorption modes of gas samples.

As shown in Figure 4.1A, radiance from the blackbody source was directed into a gas cell and then into the entrance port of the spectrometer by use of two 90° off-axis-gold coated parabolic mirrors. The mirror dimensions were 7 × 9 cm and the reflected effective focal length was 177.8 mm. A diode laser was used to align the optical path to make sure maximum radiance was received by the spectrometer.

In the laboratory data collection, ethanol (absolute 200 proof, AAPER Alcohol and Chemical Co., Shelbyville, KY) and methanol (99.8%, Spectrum, Gardena, CA) vapors were used as the analytes of interest. Liquid samples were injected by a 10 μL syringe (Hamilton, Co., Reno, NV) into a Pyrex gas cell at atmospheric pressure and allowed to vaporize at room temperature. For sample injections at or below 0.1 μL effective volume, the pure liquid was first diluted 1/10 with water and 1.0 μL of the mixture was injected into the gas cell.

The sample cell was fitted with anti-reflection coated ZnSe windows (Janos Technology, Keene, NH) with a thickness of 5 mm. The diameter of the cell was 50.8 mm and the length was 10 cm, which was used as l in the implementation of Eq. 4.3. The cell volume was estimated as 0.152 ± 0.002 L on the basis of four replicate trials of weighing the cell before and after filling it with water. The corresponding volume of water was then estimated by use of the density of water at known temperatures. The temperature of the cell was monitored by use of a Type-T thermocouple and digital thermocouple meter (Omega Engineering, Stamford, CT).

4.3.1.2 Field Data

Field data used in this work were collected by our research collaborators at the U.S. Army Edgewood Chemical Biological Center (Aberdeen Proving Ground, MD). A Brunswick FT-IR spectrometer, Model 21 (Intellitec, DeLand, FL) was used to collect the interferogram data. The spectrometer was equipped with a narrow-band MCT detector optimized to cover the 800 to 1400 cm^{-1} spectral region. The MCT detector was cooled with a closed-cycle Stirling cryogenic cooler. Interferograms consisting of 1024 points were again collected as single scans and sampled at every eight zero-crossings of the He-Ne reference laser. The maximum spectral frequency was 1974.75 cm^{-1} and the nominal spectral resolution was 8 cm^{-1}. Interferograms were collected at approximately 2.4 scans per sec.

As shown in Figure 4.1B, the spectrometer was placed about 2 m above the ground and inclined at an angle to view the exit of a portable emission stack approximately 4.6 m above the ground. The FOV of the spectrometer was restricted to 0.5° with a refractive telescope. The stack was placed 19 m away from the spectrometer. The diameter of the stack was 0.45 m. A backdrop material made of polyvinyl chloride (PVC) was located 22 m from the spectrometer. The backdrop was used to simulate a simple terrestrial background. The dimensions of the PVC material were 2.75 × 3.8 m and the measured emissivity was 0.9.

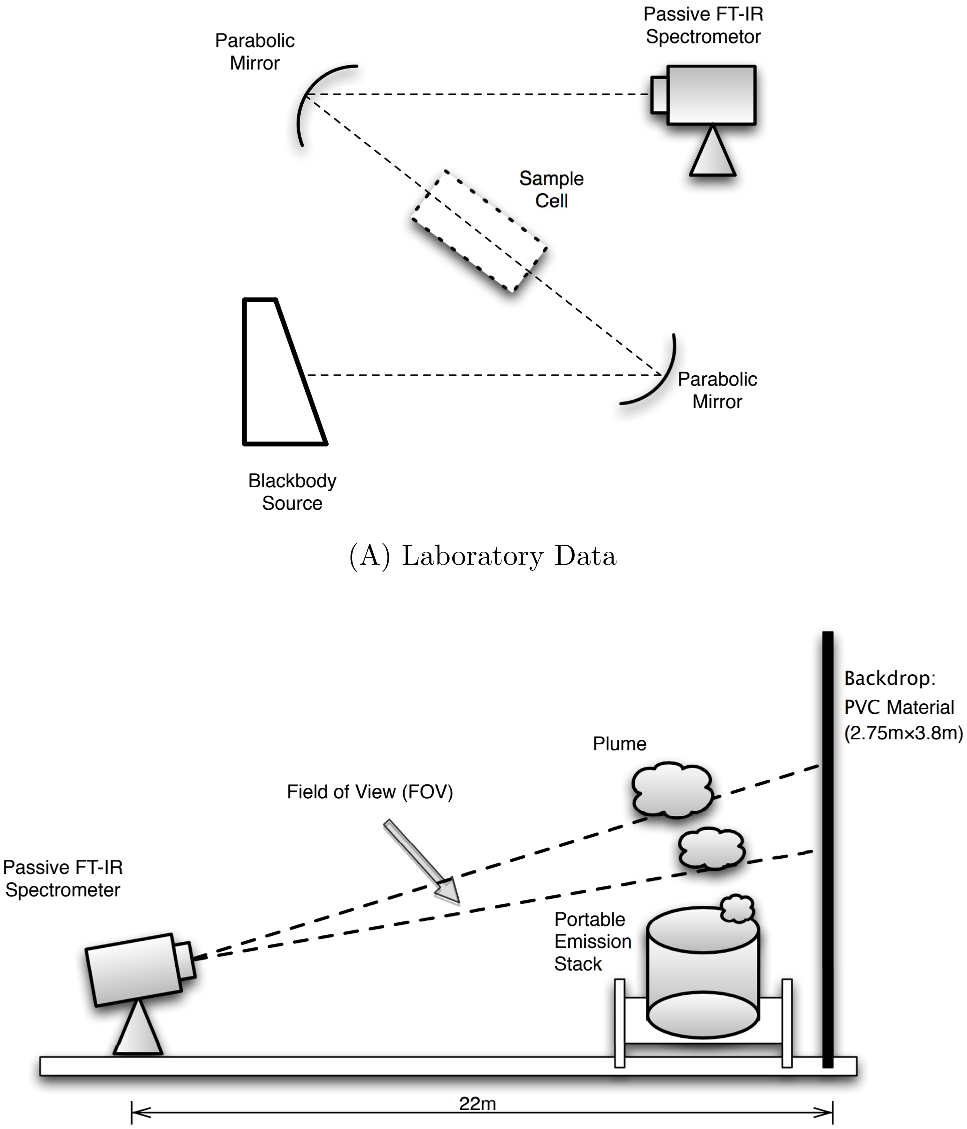
In passive remote sensing, both the ground and sky can be used as backgrounds. The chosen background depends on the view direction to the plume. The terrestrial background obtained when the spectrometer views the plume from above is more stable than the sky background when the plume is viewed from below. This measurement was configured to simulate the scenario of stack emission monitoring from above or the case in which an artificial backdrop is placed behind the stack in order to present the spectrometer with a simpler background scene.

Gas plumes were generated with a portable emission stack (Aerosurvey Inc., Manhattan, KS) that has been described previously by Chaffin.[87] The methanol was reagent grade (Tilley Chemical Company, Baltimore, MD), while the ethanol was absolute 200 proof (AAPER Alcohol and Chemical Company, Shelbyville, KY). The liquid analytes were introduced to the plume and vaporized by passing them through a hot air stream generated by a propane burner. The sample introduced to the plume was controlled by a flow meter, which had been mass-calibrated for ethanol and methanol. The concentration of each sample collection was estimated from the mass flow rates and the air velocity in the stack. The mass flow rates were monitored by measuring the weight change of the analyte in the container during the sample release. The air velocity was measured with a Pitot tube placed in the center of the stack, approximately 20 cm from the exit. The stack temperature was varied for evaluation of temperature effects. It was measured with a thermocouple placed 10 cm from the top of the stack.

4.3.2 Data Collection and Partitioning

4.3.2.1 Laboratory Data

The laboratory data were collected across three days. There were eight pure samples of ethanol and six pure samples of methanol collected. The temperature of the sample vapor was uncontrolled and allowed to equilibrate with the current room temperature (approx. 23 °C). The blackbody temperature varied between 50 and 5 °C to produce analyte spectral signals in either absorption or emission mode. Blackbody temperatures were ramped either up or down in 5° increments across the range rather than randomized. Because of instability in the collected single-beam spectra, not all samples could be studied at all temperatures.



(A) Laboratory Data

(B) Field Data

Figure 4.1. Experimental setup of data collection. (A) Laboratory experiment. The radiance was generated by the blackbody source (14 × 14 inch). Radiance, directed by the parabolic mirror (7 × 9 cm) passed through the sample cell and was directed to the spectrometer. The blackbody temperature varied between 5 to 50 °C. (B) Field experiments. The spectrometer was 2 m above the ground and 22 m away from the PVC backdrop (2.75m × 3.8 m). The FOV was restricted to 0.5° with a refractive telescope. The emission stack was placed in front of the backdrop and 19 m to the spectrometer. The plume was generated at three temperatures of 150, 175, and 200 °C.

With each prepared ethanol or methanol sample, the data collection protocol employed three experimental configurations: (1) collection of open-beam data (i.e., no sample cell present), (2) collection of blank data in which the freshly evacuated cell was placed in the optical path, and (3) acquisition of sample cell data for each prepared ethanol or methanol sample. With each experimental configuration, blackbody temperatures were sampled at various steps over the range of 5 to 50 °C. At each background temperature level, 100 interferograms were collected.

After injecting each sample into the gas cell, the cell was allowed to equilibrate for 1-3 h. After equilibration, the cell was placed in a conventional laboratory FT-IR spectrometer (Bruker Vertex 70, Bruker Optics, Billerica, MA or Nicolet 6700, Thermo-Nicolet, Corp., Madison, WI) and an absorbance spectrum was obtained at a nominal resolution of 4 cm^{-1}. For six of the 14 total samples, a second replicate absorbance spectrum was also measured to assess the concentration stability within the gas cell.

Sample concentrations were estimated in two ways. Using tabulated densities for ethanol and methanol of 0.7893 and 0.7914 g/cm^3,[88] the delivered volumes were converted to mass and then to moles. Air pressure and temperature measurements in the laboratory at the time of the data collection were then used to estimate the number of moles of air in the cell assuming ideal gas behavior. The path averaged concentration in ppm-m was estimated on the basis of the cell path length of 0.1 m and the ratio of moles of analyte to moles of air.

Concentrations were also estimated by use of the collected absorbance spectra.

In each spectrum, the range of 700 -1700 cm^{-1} was fitted by multiple linear regression to a pure-component spectrum of the analyte at 1.0 ppm-m taken from the Pacific Northwest National Laboratory (PNNL) quantitative IR vapor phase library[89]. The fit also included a 2$^{\text{nd}}$-order polynomial baseline term in each case and a PNNL spectrum of water for the samples based on mixtures of water and analyte. The PNNL spectra used were all collected at 25 °C and a resolution of $\sim$0.1 cm^{-1}. Before fitting to the experimental spectra, the PNNL spectra were deresolved to match the resolution of the Thermo and Bruker spectrometers. The deresolving procedure involved the convolution of the PNNL spectrum with a boxcar windowing function, with the width of the window optimized to maximize the quality of fit to the experimental spectra. For the data collected with the Thermo and Bruker spectrometers, respectively, windowing functions with widths of 4.16 cm^{-1} and either 6.15 cm^{-1} or 6.33 cm^{-1} produced the best fits.

Comparison of the results obtained with the two concentration estimation procedures revealed results within 1.3 – 6.5% for ethanol samples and 13.9 – 17.0% for methanol samples for which the injection volume was at least 0.75 μL. Greater differences were noted for smaller injection volumes. Concentrations based on the absorbance spectra were consistently less than those based on direct calculation. This was assumed to be an indication of incomplete vaporization of the liquid sample in the gas cell. For this reason, coupled with uncertainties in the injection volumes, the concentrations based on the absorbance spectra were used in subsequent calculations. For the cases in which replicate absorbance spectra were collected, differences in concentration estimates across replicates were always less than 1%. This indicated that stable concentrations were achievable in the gas cell and provided confidence that the concentration estimates were applicable to the passive IR spectra.

The maximum concentration of ethanol was 2432.5 ppm-m with the minimum at 44.9 ppm-m. The concentration range of methanol was 37.4 to 1181.9 ppm-m. A summary of the collected data is provided in Table 4.1.

The 100 interferograms of the blackbody with and without the empty sample cell in the light path were used to estimate the instrumental parameters (responsivity and self emission), which are important in synthesis of the spectrum.

In obtaining replicates of sample spectra, different interferogram co-addition levels were compared. With 100 interferograms, co-adding 50 inteferograms produces two replicates for each sample, and a co-addition level of 30 will result in three replicates. The spectral noise level is usually used to evaluate the quality of the data. Random noise should cancel according to the square root of the number of co-added scans. The noise level can be computed from the ratio among the replicate spectra by calculating the absorbance values from the transmittance as:

$$A_i = -\log \frac{P_{rep1,i}}{P_{rep2,i}} \tag{4.6}$$

where A_i denotes the absorbance value, $P_{rep1,i}, P_{rep2,i}$ are the intensities of the single-beam spectra corresponding to the replicates, and i represents the spectral point (i.e., wavenumber) at which the value is used. The resulting absorbance spectrum is termed a '100% line'. For n replicates, $n!/(2!(n-2)!)$ '100% lines ' can be obtained by calculating all possible combinations from the replicates. Theoretically, the 100% line should reflect random noise about a flat baseline at 0.0 absorbance units (AU). However, instrumental variation or environmental changes during the data collection can produce a non-zero baseline (i.e., a systematic component in the 100% line). For this reason, the baseline can be modeled by a polynomial function and the random noise about this baseline can be used as a measure of the spectral noise level.

In the work reported here, the noise was estimated by the root-mean-squared (RMS) error of the deviations between a fitted quadratic baseline function and the computed 100% lines. Equation 4.7 details this calculation.

$$\text{RMS} = \sqrt{\frac{\sum_{k=1}^{n} d_k^2}{n - (df + 1)}} \tag{4.7}$$

In the equation, d_k denotes the deviations from the fitted baseline model, n represents the number of spectral points involved in the calculation, df specifies the degrees of freedom associated with the model coefficients (2 for a quadratic model), and '+1' is for the intercept term in the baseline model.

In evaluating the differences among the co-addition levels, the spectral range from 950 to 850 cm^{-1} was used to calculate the RMS noise values. Figure 4.2 shows the RMS values of replicate spectra over the three days of data collection. It can be observed that, for a large amount of samples, the noise level was lower at the higher co-addition level. Samples where this was not the case included some systematic variation across the replicates. Therefore, in the further data analysis, two replicates

Table 4.1. Sampling Profile of Laboratory Data Collection

Analyte	Background Temp.(°C)	No. of Sample
Ethanol	50/45/35/25 40/30/20/15/10/5	3[a] 8
Methanol	40/30/20/15/10/5	6

[a] These represent a subset of the 8 total ethanol samples.

based on 50 co-added interferograms were used. Because no replicates were needed for background spectra, all interferograms were co-added to obtain the spectra of the blackbody radiance at the different temperatures.

4.3.2.2 Field Data

The field data were collected over six days with the stack temperatures varied among 150, 175 and 200 °C. The ambient temperatures over the six days ranged over 14 to 30 °C. There were 225 samples released in total. Approximately 500 interferograms were collected for each sample. A sample was either a pure-component release of methanol or ethanol, or a mixture of ethanol and methanol at a given stack temperature. Table 4.2 lists the sampling details in the data collection. The concentration range of the pure ethanol releases was 28 - 284 ppm-m, while the corresponding range of methanol was 7 - 284 ppm-m.

A uniform experimental design[90] was used in designing the targeted mixture concentrations. For mixture releases, the maximum concentration was 276/270 ppm-m with the minimum at 21/135 ppm-m in terms of the ethanol/methanol ratio. Figure 4.3 depicts a plot of the concentration distribution of methanol and ethanol. The

Table 4.2. Sampling Profile of Field Data Collection

Day	Analyte	T_{Stack} (°C)	No. of Sample
1	Ethanol	175 200 150	12 12 12
2	Ethanol	175	29
3	Ethanol	200 150	30 27
4	Methanol	175 200 150	11 12 12
5	Methanol/Ethanol Methanol	175	34 11
6	Methanol	175	23

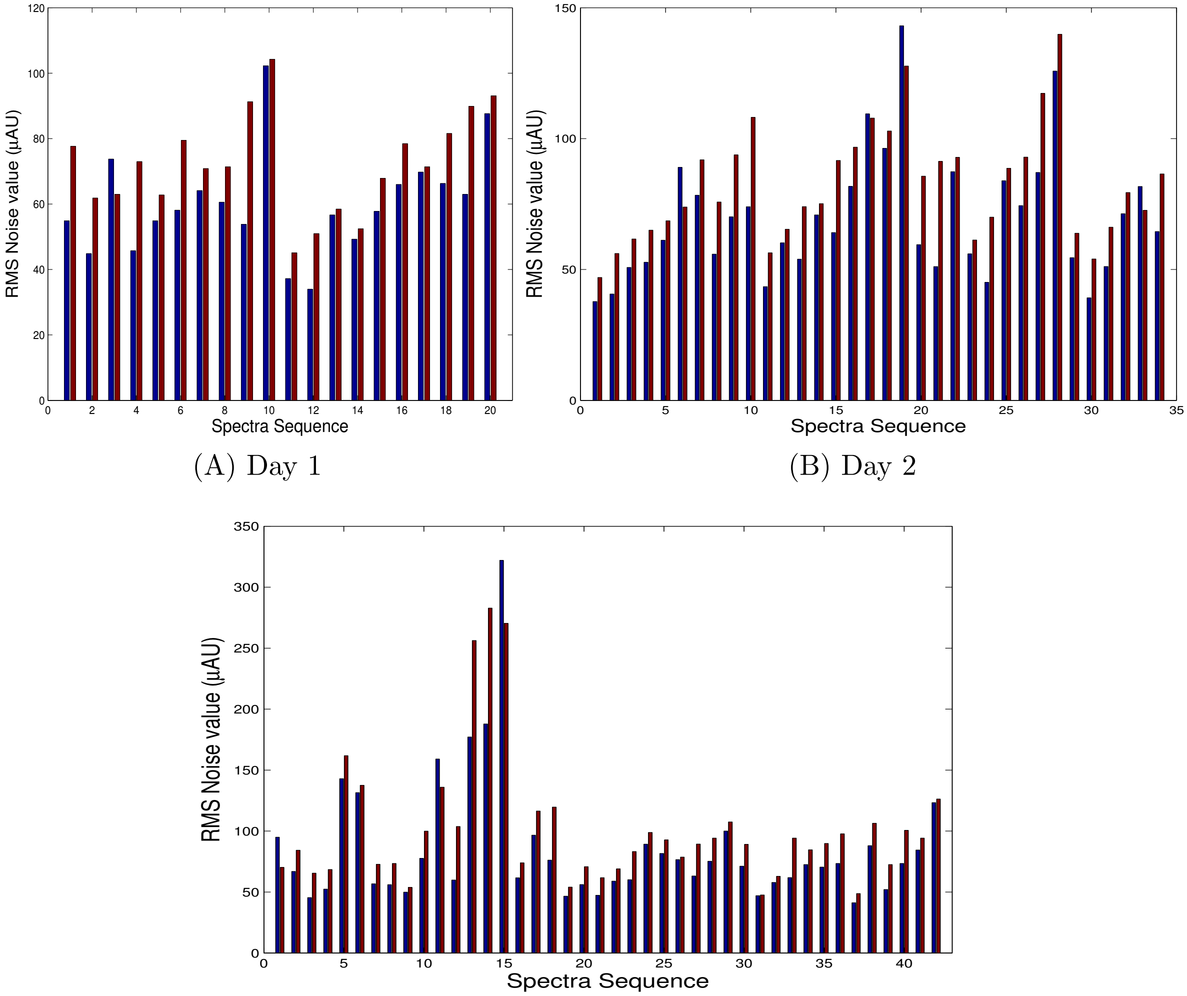
random scattering indicates a low correlation between the two analytes.

For both pure and mixture samples, the path-averaged concentration was estimated on the basis of the stack exit diameter of 0.45 m and the ratio of the sample vapor emission rate to the total air flow rate:

$$\text{Concentration(ppm} - \text{m)} = \frac{\text{Vapor Emission(ft}^3\text{/min)}}{\text{Total Air Flow(ft}^3\text{/min)}} \times 0.45\text{(Stack Exit Diameter(m))} \tag{4.8}$$

Besides the sample data collection, interferograms of an external blackbody source at different temperatures were also collected on each day. This allowed the calculation of the instrument responsivity and the self-emission function. The temperature of the blackbody source varied between 10 and 115 °C.

Similar to the laboratory data, two replicate spectra were obtained for each sample spectrum. An interferogram co-addition level at 200 was selected to assemble the sample spectra for further data analysis.



(A) Day 1

(B) Day 2

(C) Day 3

Figure 4.2. Bar plots of RMS noise values for groups of replicate spectra in the laboratory data collection. The blue bars represent the RMS noise values obtained from co-adding 50 interferograms, which results in two replicate spectra for each sample. The red bars denote the noise values obtained when the co-addition level was 30 and three replicates were obtained.

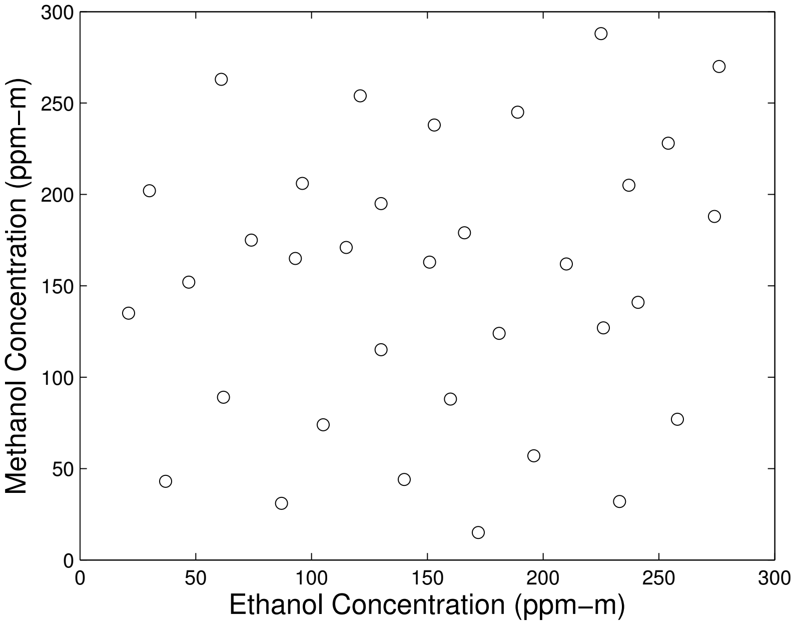
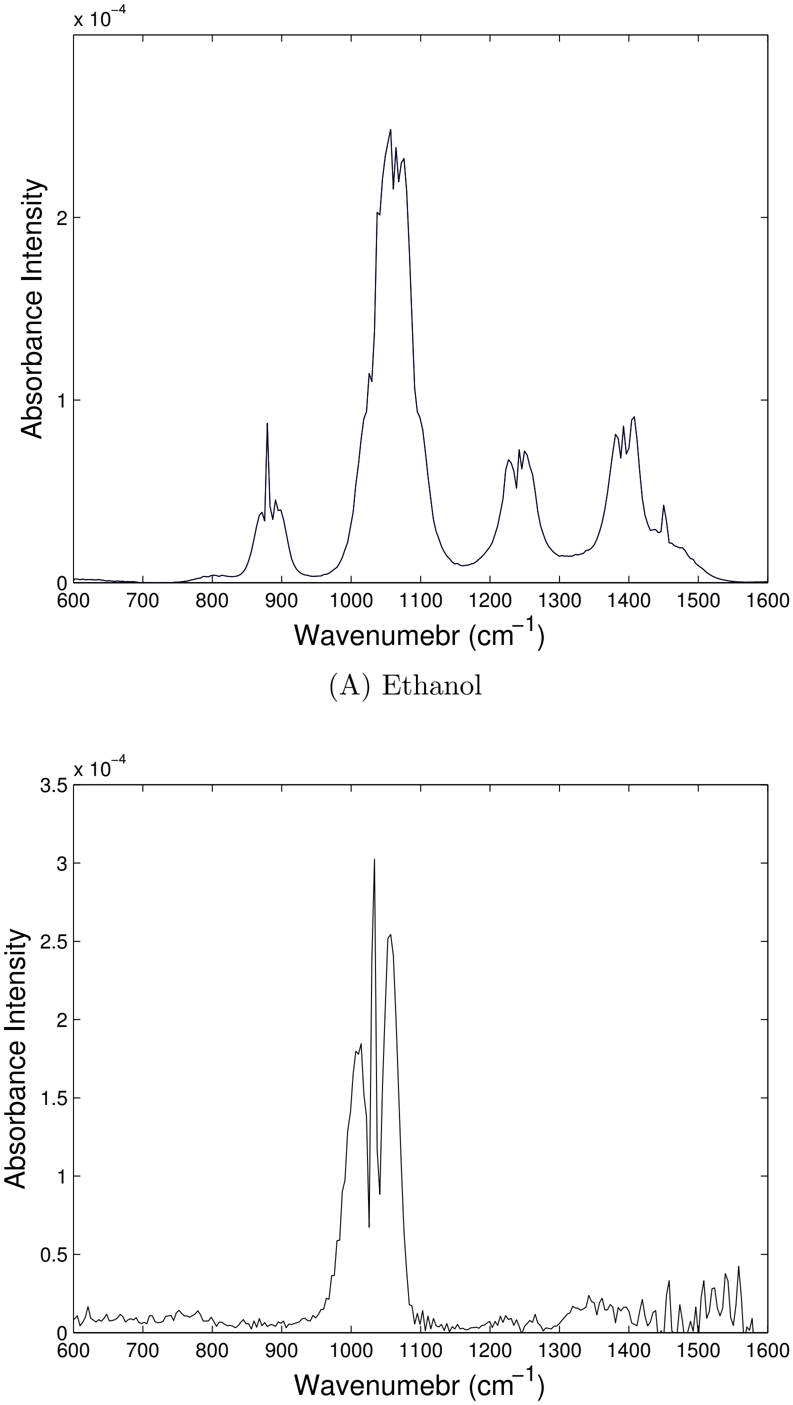


Figure 4.3. Scatter plot of methanol concentration vs. ethanol concentration for mixture sample release on Day 5. The circles are randomly scattered, which indicates that the correlations between methanol and ethanol concentrations are low.



(A) Ethanol

(B) Methanol

Figure 4.4. Pure-component spectra of ethanol (A) and methanol (B). The effective burden was 1 ppm-m. The point spacing was reduced from 0.06 cm^{-1} to 4 cm^{-1}.

4.3.3 Computation

All computation of laboratory data and field data was performed under MATLAB (version 7.4 , The MathWorks, Inc., Natick, MA) running on a Dell Precision 670 workstation (Dell Computer Corp., Austin, TX) operating under Red Hat Enterprise Linux WS (Red Hat, Inc., Raleigh, NC).

4.4 Results and Discussion

4.4.1 Description of Simulation Procedure

The basic premise of the quantitative analysis procedure employed here was to apply Eqs. 4.1, 4.2 and 4.3 to synthesize calibration data, to use the resulting calibration data to build PLS models for predicting analyte concentrations, and to apply the resulting models to predict 'unknown' concentrations in the experimental data collected in the laboratory and field. Through this procedure, new calibration models could be constructed as often as needed to reflect changes in experimental conditions.

The detector responsivity (r) and instrument self emission function (L_e) terms in Eq. 4.3 were estimated from the hot and cold blackbody measurements performed on each day. With the assumption of linear detector response, any two temperatures would be sufficient to determine the instrumental terms. In practice, the blackbody source temperatures were selected in the range which would provide a close match to the single-beam intensities of the sample spectra.

Tables 4.3a and 4.3b list the temperatures of the blackbody sources used in

Table 4.3. Blackbody Source Temperatures Used in Estimating r and L_e

(a) Laboratory Data

Targeted Background Temperature (°C)	Blackbody Source Temperature Warm (°C)	Cold (°C)
55 - 45	50	45
45 - 40	45	40
40 - 35	40	35
35 - 30	35	40
30 - 25	30	35
25 - 20	25	20
20 - 15	20	25
15 - 10	15	10
10 - 0	10	5

(b) Field Data

Day	Blackbody Source Temperature Warm (°C)	Cold (°C)
1 2 3	45	30
4	50	35
5	35	10
6	30	15

calculating the instrumental responsivities and self-emission functions for the laboratory and field data, respectively. To determine the L_x term in Eqs. 4.1 and 4.2, the spectral radiances of the background and analyte cloud are needed, as well as the analyte transmittance. The analyte transmittance is determined by the absorptivity and the path-averaged concentration. Based on the assumption of unit emissivity of the background and the use of Planck's function to compute L_t associated with the emission of the analyte cloud in Eq. 4.1, the spectral radiance values are determined by the background and sample temperatures. Therefore, in order to generate the synthetic calibration spectra, the background temperature, T_{bkg},and analyte temperature, T_t, are required.

4.4.2 Estimation of Background and Analyte Temperatures

For both the laboratory and field data, attempts were made to estimate background and analyte temperatures during the data collection through experimental measurements. It was found, however, that in some cases these temperatures did not match the apparent temperatures reflected in the collected single-beam spectral intensities. This was particularly problematic in the field data where analyte temperatures were measured inside the stack but the spectral measurements acquired radiance from above the stack exit. Similarly, the ambient air temperature was used to estimate the temperature of the PVC backdrop that served as the target for the spectral background, not considering the potential effect of direct solar heating of the material. Consequently, the collected single-beam spectra were used to estimate the apparent temperatures as described separately below for the laboratory and field data.

4.4.2.1 Laboratory Data

In the data collection, a controlled-temperature blackbody was used to define the background radiance. However, this presupposed that only radiance from the blackbody was in the FOV of the spectrometer. As shown in the diagram of the experimental setup in Figure 4.1, this required both a stable blackbody temperature and precise alignment of the optics. Inspection of the data revealed some inconsistencies in the single-beam intensities as they related to the temperature settings of the blackbody source. Hence, it was necessary to verify the background temperature before the sample quantification through a spectral fitting procedure. For the laboratory data, if the fitted background temperature deviated significantly from the expected value, the data at that background temperature were not used in further calculations.

To estimate the background temperatures, the collected sample spectra were studied individually. The strategy was to match the intensity of the synthesized spectra with the collected spectra by varying the temperature inputs. The goal was to seek the temperature which provided the minimum summed intensity difference between the simulated and collected spectra as defined by h in Eq. 4.9

$$h = \sum_{i=1}^{n} |I_{Simulated,i} - I_{Experimental,i}| \tag{4.9}$$

In the equation, $I_{Simulated,i}$ and $I_{Experimental,i}$, i represent single-beam spectral intensities for simulated and experimental data, respectively, at point i across the range of n wavenumber points used. To estimate the background temperature, as shown in the library absorbance spectra in Figure 4.4, there are wavenumber regions where no spectral features of the analyte can be observed. Table 4.4 lists the spectral ranges used to estimate the background temperatures for the methanol and ethanol data.

The gas sample in the laboratory setup was obtained from the vaporized liquid under room temperature. For the work performed with the laboratory data, the analyte temperature was assumed to be equal to the measured room temperature.

4.4.2.2 Field Data

As noted previously, significant variation was observed in the field data that appeared inconsistent with the measured temperatures. Because the field data were collected outdoors in an uncontrolled environment, they were subject to meteorological variation that was not captured in the temperature measurements. Issues of optical alignment are also applicable to the field data.

To estimate the temperatures, sample spectra were again studied. As presented in Table 4.2, data were studied in subgroups according to their collection days and the stack temperatures. On each day, the first sample released at a certain stack temperature was studied for temperature estimation. To obtain replicate spectra for the estimation procedure, the 500 interferograms were co-added in groups of 10 to produce 50 replicates. The spectral region and procedure used in estimating the background temperature was the same as in the study of the laboratory data.

In estimating the analyte temperature, the region which contained the strongest analyte feature was selected. Table 4.4 lists the spectral ranges used to estimate the analyte temperature for the pure and mixture samples.

During the data collection in the field, the analyte temperature was found to be unstable due to changes in environmental conditions. Another complicating factor was that the analyte concentration was calculated based on the mass flow rate of the analyte and the air velocity. Any variation in mass flow rate or air velocity could cause fluctuations in the sample release, thereby affecting the actual observed concentration. However, during a given sample collection of 500 interferograms ($<$

5 min), the concentration was assumed to be constant. While both temperature and concentration variation can cause changes in the observed single-beam spectral intensity, with the assumption of a stable concentration release, all fluctuations were attributed to variation in the the analyte temperature.

In the analyte temperature study, an analyte temperature higher than the stack temperature was sometimes obtained. In reality, however, gases coming out of the stack cannot be hotter than the stack temperature. This result could arise because a sample was released with a higher concentration than the calculated concentration value or because of a lower estimation of the background temperature. In this case, the estimated temperature cannot be used. Similarly, a lower concentration release will result in a lower analyte temperature estimation. However, for the open-air experiment, there is no reference method to confirm the concentrations except the theoretical calculation. Therefore, a diagnosis of sample release stability was required. This was done by forcing the analyte temperature to fall in a particular range, not higher than the stack temperature or lower than a certain value. Table 4.5 lists the criterion by which spectra were discarded in the estimation of the analyte temperature at different stack temperatures. For a sample release, if more than 30 % of the spectra in the release failed to meet the criterion, the sample was considered to be an unstable release. In this case, the next sample release was studied to estimate the temperature.

Table 4.4. Spectral Segment Used in Temperature Estimation

Sample Sepctra (cm^{-1})	Background	Analyte
Ethanol	960 – 940	1085 – 1045
Methaol	955 – 935	1055 – 1015
Mixture	960 – 935	1085 – 1015

Table 4.5. Criterion of T_{Analyte} Range of Field Data

Stack Temperature (°C)	Minimum (°C)	Maximum (°C)
150	70	150
175	80	175
200	100	200

4.4.3 Generation of Calibration Models

After obtaining the background temperature, to predict the concentration of each laboratory spectrum, a synthetic calibration data set with 300 spectra was generated. The plug-in analyte temperature was the current room temperature, while the background temperature was randomly selected from a normal distribution with the mean at the value obtained previously and a standard deviation of 0.5 °C. The path-averaged concentrations were randomly selected from the range of 50 to 2000 ppm-m.

For the field data, all spectra in the calibration set were given a fixed analyte temperature, which was taken as the median of the estimated temperatures across the group of replicate spectra used in the temperature estimation. Values discarded according to the criteria discussed above were not included in the calculation of the median. The background temperature in each spectrum was randomly selected from a normal distribution with the same mean and standard deviation values obtained in the sample study. The path-averaged concentration was randomly selected in the range of 5 to 300 ppm-m. In modeling the mixture samples, the random concentrations of methanol and ethanol were generated separately. The size of the calibration set was 200 spectra.

A partial least-squares (PLS) model was built with the synthetic calibration data. The spectral range was selected according to the absorption bands of the analyte. For the ethanol model, the spectral range was 1150 to 950 cm^{-1}, while that of methanol was 1100 – 950 cm^{-1}. The number of latent variables was chosen as two for single-component samples, incorporating one factor to model the background and one to model the analyte. For mixture samples, one more factor was added to the model to account for the interference. The spectral range remained the same as in the single-analyte model.

4.4.4 Simulation Results

Simulated spectra with the corresponding laboratory or field spectra are plotted in Figure 4.5. Good agreement can be observed between the synthetic and experimentally collected spectra. This visual comparison provides validation for the radiance model and simulation procedures adopted in this research.

For the laboratory data, Figure 4.6 shows the estimated background temperature for each replicate spectrum at different blackbody temperatures. At blackbody temperatures of 20 and 25 °C, which were very close to the analyte temperature around 23 °C, the spectral features of either emission or absorption were extremely weak and poor quantitative modeling results were obtained. Therefore, results at these temperatures are not shown. Inspection of Figure 4.6 reveals that the estimated blackbody temperatures were different in a number of cases from the theoretical blackbody temperatures corresponding to the source setting. As noted previously, these deviations may have been in part due to imperfect alignment of the optical components. Because of the uncertainty associated with these cases, if the obtained value for the background temperature was ± 2 °C different from the theoretical value, the spectrum was eliminated from further data analysis.

Tables 4.6a and 4.6b list the values of the standard error of prediction (SEP) obtained by using the simulated calibration spectra for ethanol and methanol, respectively. The SEP values are given in absolute terms in units of ppm-m, as well as the median percentages relative to the reference concentration. The percentage relative error is calculated by Eq. 4.10, where c_{Pred} and c_{Actual} are the estimated and reference concentrations, respectively.

$$\text{Relative Error\%} = \frac{|c_{\text{Pred}} - c_{\text{Actual}}|}{c_{\text{Actual}}} \times 100 \tag{4.10}$$

The corresponding correlation plots are shown in Figure 4.7 and 4.8 for ethanol and methanol, respectively. The correlation plots at 50 °C, 45 °C, and 35 °C are not shown. Because only three samples were collected at those temperatures, there were not enough spectra to obtain a representative result.

It can be observed that at lower concentrations (e.g., below 1000 ppm-m), the estimated concentrations correlate better with the reference values than at larger concentration levels. The SEP value indicates that, the higher the temperature difference between the blackbody and sample, for example at background temperatures of 40 °C and 5°C, the better the prediction performance. This conclusion is expected on the basis of how temperature contributes to the spectral intensities observed. According to Planck's function, the spectral intensity is primarily dependent on the temperature. In a larger temperature difference scenario, analyte and background information can be distinguished better. In addition, the methanol predictions were generally more accurate than those for ethanol. This could be because the spectral feature of methanol at around 1040 cm^{-1} is sharper than ethanol.

From the variations in estimated background temperatures and the RMS noise values, the data were not very stable among samples and between sample replicates. In the estimation of the background temperatures, single-beam spectra were used. The inconsistency of the observed radiance intensities at different wavenumbers can potentially induce bias in the background temperature estimation and thereby affect the quantification of the sample. For example, Figure 4.9 shows the replicate spectra for one of the ethanol samples collected on day 2. It can be observed that for the replicates of a sample, the spectral intensities between 700 and 800 cm^{-1} are consistent. However, the signal starts to deviate from 800 to 1400 cm^{-1}. Meanwhile, the temperature differences between sample and background in the laboratory data are low. The maximum temperature difference is around 18 °C. The analyte information superimposed on the background spectra is relatively weak for these data.

In terms of the field data results, Table 4.7 lists the simulation results in terms of the estimated background and analyte temperatures and the corresponding SEC and SEP values at different stack temperatures on each day. Also listed are the median % relative errors. The correlations between predicted analyte concentrations and reference values are plotted in Figures 4.10, 4.11, and 4.12 for the ethanol, methanol, and mixture releases, respectively.

The calibration errors in this study were much lower than the prediction errors. The simulated calibration data are very stable, because no variation information is added manually except the background temperature fluctuation. However, collected outdoors, the gas releases are subject to be influenced by changes in environmental conditions and the data collection can be affected by instrumental drift. As concluded with the laboratory data results, the prediction performance also depends on the temperature difference between background and sample. In the field data collection, the background temperatures were much lower than the analyte temperatures. The larger difference between analyte and background can potentially lower the effect of inaccurate background temperature estimation. Meanwhile, from a hotter stack, a higher analyte temperature can be obtained, and thereby more radiance can be emitted by the analyte. Consequently, the analyte signal is stronger and better prediction results are made possible. For the mixture releases, the ethanol predictions were better than those for methanol according to the SEP values. However, from the correlation plots, bias in the predictions can be observed and methanol was mostly under estimated. Such results could be caused by interference from the presence of ethanol.

In comparison of the laboratory and the field results, similar conclusions can be obtained in terms of the temperature and analyte. However, the SEP values of the laboratory data are higher than those of the field data, primarily because of the larger concentration range studied.

4.5 Conclusions

In this chapter, a simulation method for calibration in passive FT-IR remote sensing was successfully developed. Obtaining calibration information in remote sensing measurements is challenging, because the background is unstable and data collection is labor intensive and expensive. The simulation strategy investigated in this study can lower the cost in collecting calibration data and mimic the prediction data by measuring or estimating key parameters that influence the data.

The laboratory and field data were studied with similar simulation procedures. Because of the experimental setup, both emission and absorption spectral features of the analytes were obtained from the laboratory data, while only the emission mode was available for the field data. The prediction performance was better when higher temperature differences between sample and background were obtained.

Based on the temperature information obtained from the collected sample or the experimental conditions, the calibration data can be synthesized for future prediction. Because the calibration model is built with the single-beam spectra, no background spectra are needed in this method. In this work, the only required background data collection was the measurement of spectra at two blackbody temperatures for the purpose of calculating the instrumental response and self-emission parameters needed for the spectral synthesis.

However, to make this method work practically in the field, a stable sample release is required to obtain the temperature information of both the background and sample. Similarly, if the environmental conditions change, additional data collection is required. But as long as the temperature information is obtained, calibration data can be generated very quickly. This makes it possible to have a large block of calibration data without actual data collection. Finally, it is important to put the prediction performance of the described methodology into proper context. Relative errors of 20% might be unacceptable in a laboratory analysis but could be extremely informative and valuable in a field monitoring application where currently there is no simple method to obtain quantitative information from a released gas. Furthermore, the methodology described here could be adapted to applications where instead of a precise concentration estimate, a simple yes/no answer is desired regarding whether an emission exceeds a threshold limit that signals a gross error or equipment failure.

Table 4.6. Prediction Results for Laboratory Data

(a) Ethanol

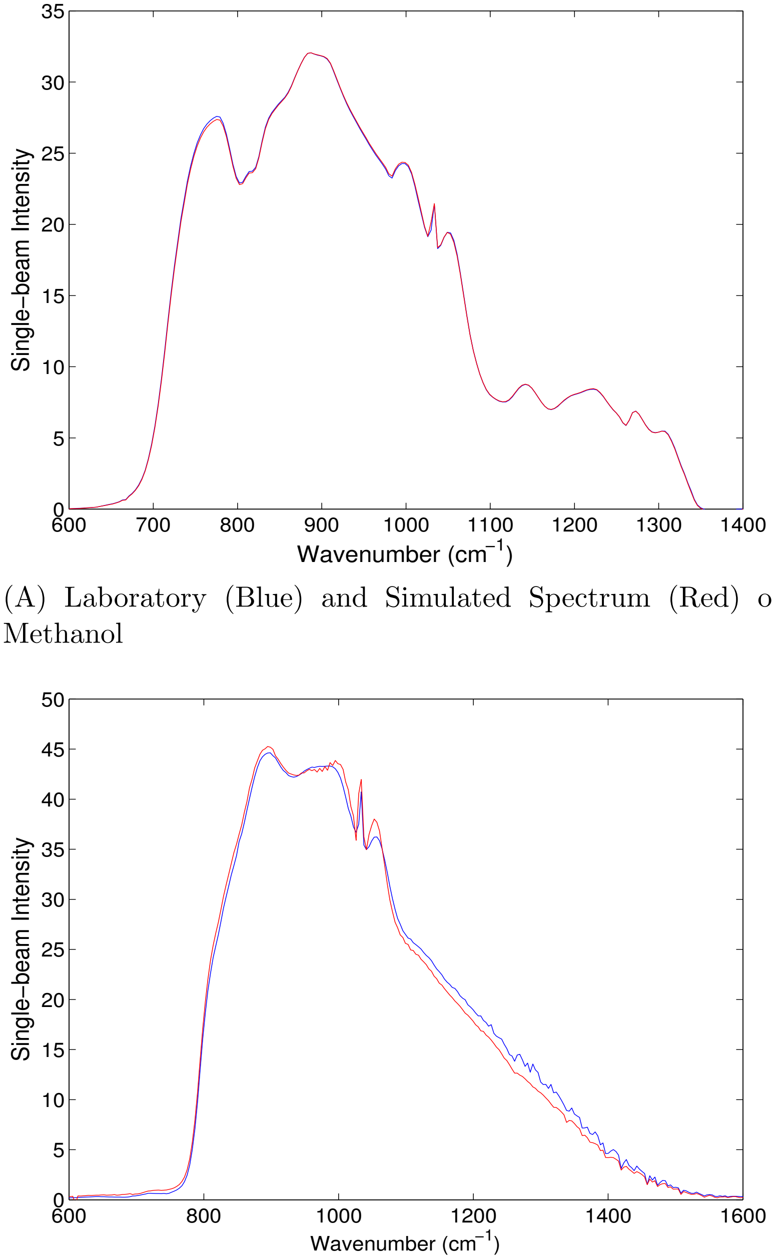
$T_{\text{Blackbody}}$	SEP of Ethanol	Relative Error	No. of Spectra
(°C)	(ppm-m)	(%, Median)	(Used/Original)
40	52	9.1	12/16
30	117	22.4	10/16
15	234	18.7	8/16
10	257	17.7	10/16
5	172	17.6	11/16

(b) Methanol

$T_{\text{Blackbody}}$	SEP of Methanol	Relative Error	No. of Spectra
(°C)	(ppm-m)	(%, Median)	(Used/Original)
40	27	60.9	8/12
30	130	54.7	9/12
15	147	46.4	10/12
10	101	32.5	10/12
5	98	18.1	10/12

Table 4.7. Prediction Results for Field Data

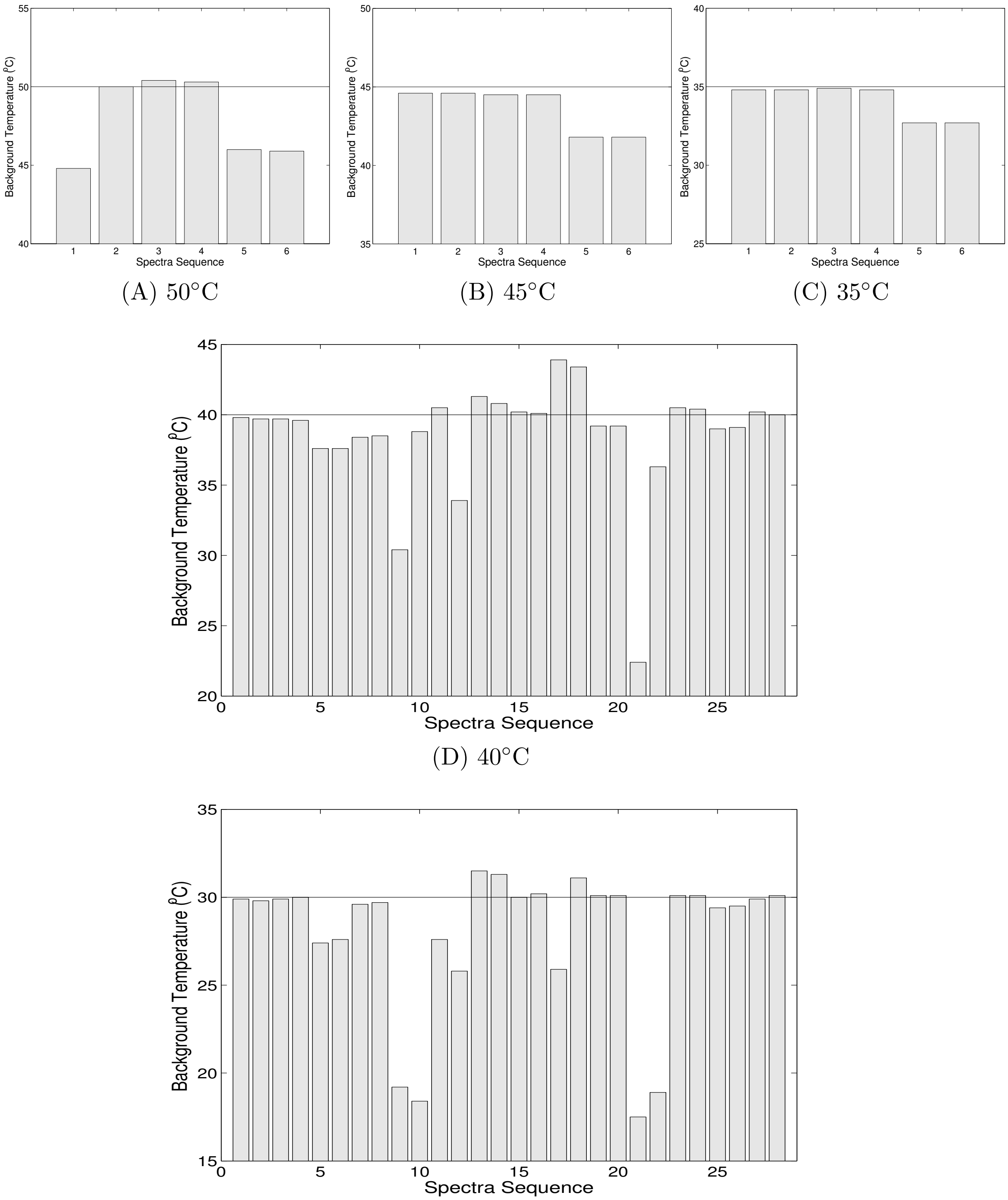
Day	Analyte	T_{Stack}	$T_{\text{Background}}$	T_{Analyte}	SEP	Relative Error
		(°C)	(°C)	(°C)	(ppm-m)	(% Median)
1	Ethanol	175	33.6 ± 4.3	120	49.3	22.1
		200	43.1 ± 0.2	117	28.0	11.7
		150	27.0 ± 0.3	120	31.3	31.3
2	Ethanol	175	32.8 ± 1.6	111	49.4	23.2
3	Ethanol	200	24.1 ± 0.1	119	23.3	13.8
		150	21.5 ± 0.3	102	46.5	22.5
4	Methanol	175	16.6 ± 0.3	154	24.4	15.8
		200	10.3 ± 0.2	162	17.8	12.5
		150	8.5 ± 0.2	131	77.0	30.7
5	Ethanol/Methanol	175	17.1 ± 0.3	164	27.5/41.0	8.6/18.0
	Methanol		9.4 ± 0.5	140	21.1	8.9
6	Methanol	175	40.6 ± 0.7	144	18.3	8.9



(A) Laboratory (Blue) and Simulated Spectrum (Red) of Methanol

(B) Field (Blue) and Simulated Spectrum (Red) of Methanol

Figure 4.5. Example of experimental (blue) and simulated (red) spectra. (A) Laboratory spectra of methanol. The blackbody temperature was 5 °C. The concentration was 880 ppm-m. (B) Field methanol spectra. The stack temperature was 175°C. The concentration was 180 ppm-m.



(A) 50°C

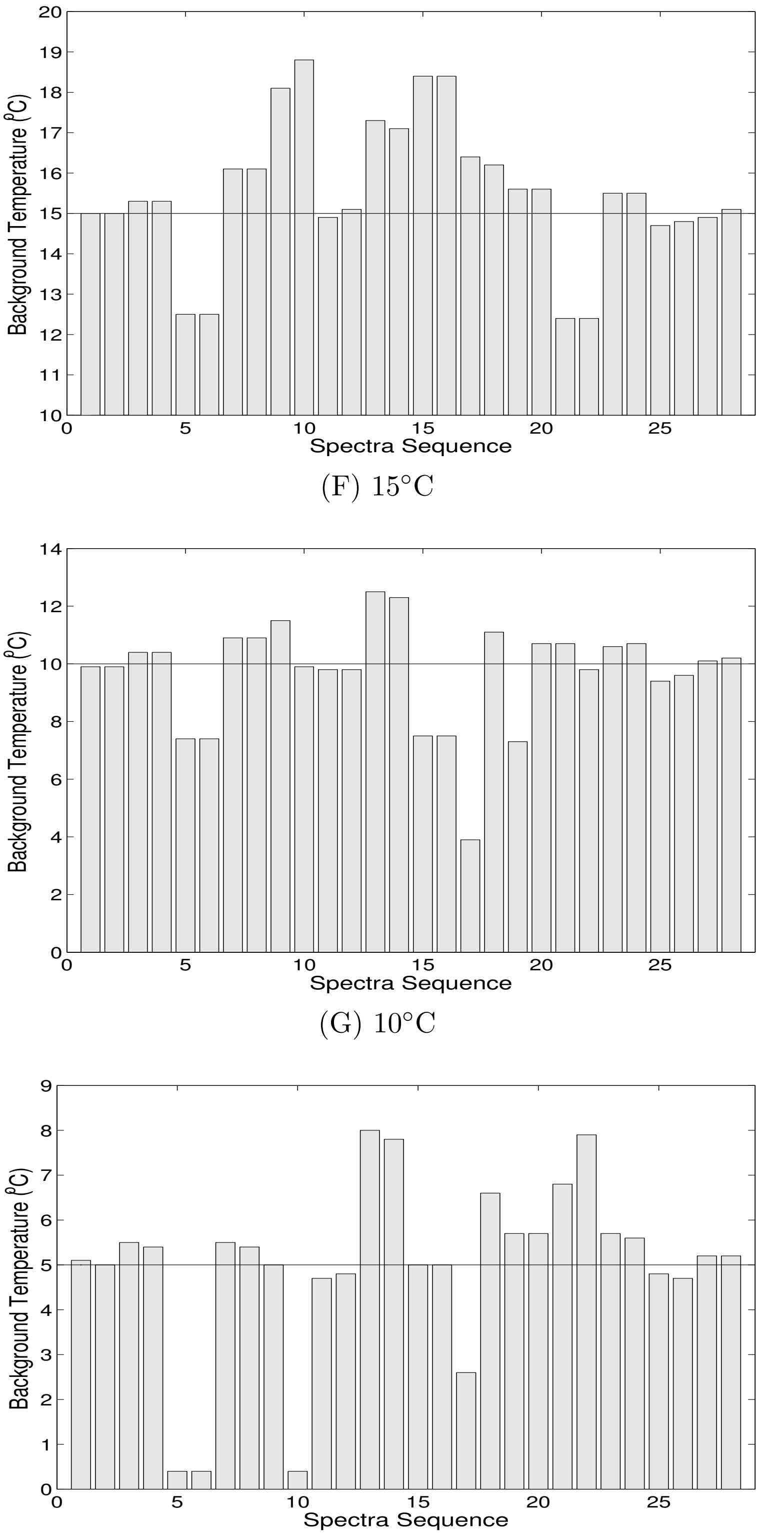
(B) 45°C

(C) 35°C

(D) 40°C

(E) 30°C

Figure 4.6. Estimated background temperatures for laboratory data at each blackbody temperature. Each bar shows the estimated background temperature based on the obtained replicate spectrum. The horizontal line across the bars shows the corresponding blackbody temperature.

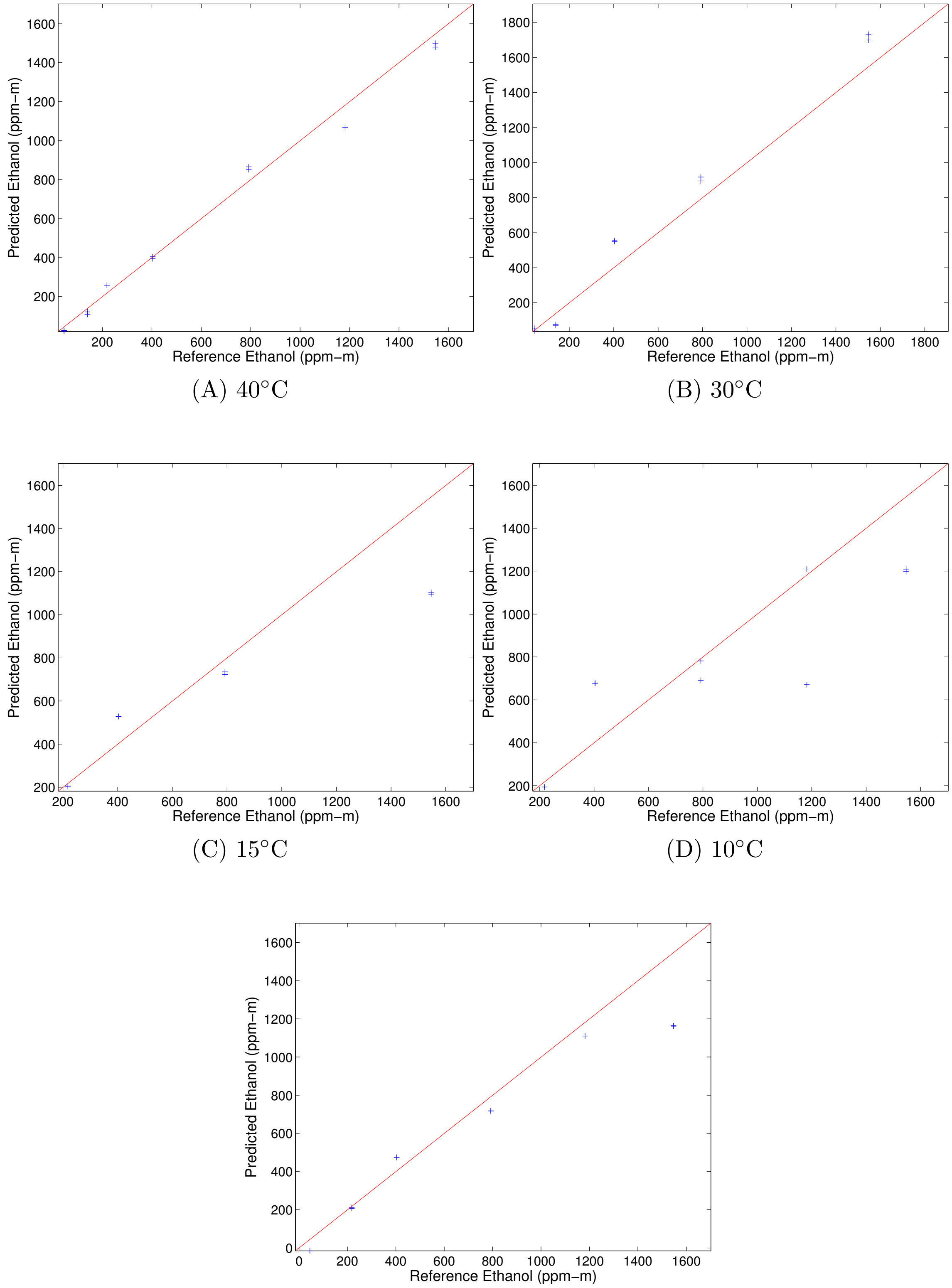


(F) 15°C

(G) 10°C

(H) 5°C

Figure 4.6. Estimated background temperatures for laboratory data at each blackbody temperature. Each bar shows the estimated background temperature based on the obtained replicate spectrum. The horizontal line across the bars shows the corresponding blackbody temperature



(A) 40°C

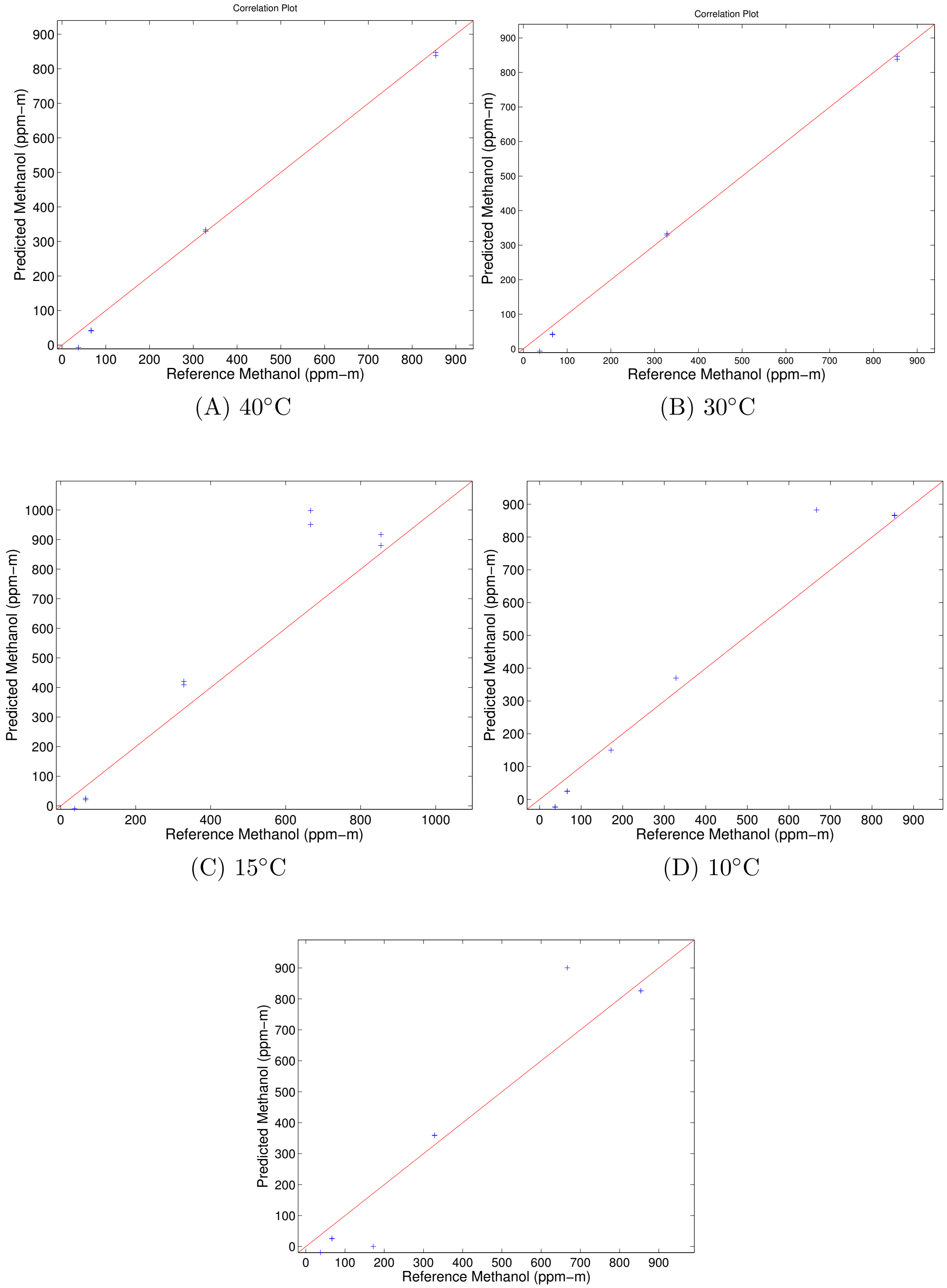
(B) 30°C

(C) 15°C

(D) 10°C

(E) 5°C

Figure 4.7. Correlation plots of predicted and reference concentrations for ethanol predictions of laboratory spectra at each blackbody temperature with the PLS model. The spectral range was 1150 – 950 cm^{-1} with two PLS factors.



(A) 40°C

(B) 30°C

(C) 15°C

(D) 10°C

(E) 5°C

Figure 4.8. Correlation plots of predicted and reference concentrations for methanol predictions of laboratory spectra at each blackbody temperature with the PLS model. The spectral range was 1150 – 950 cm^{-1} with two PLS factors.

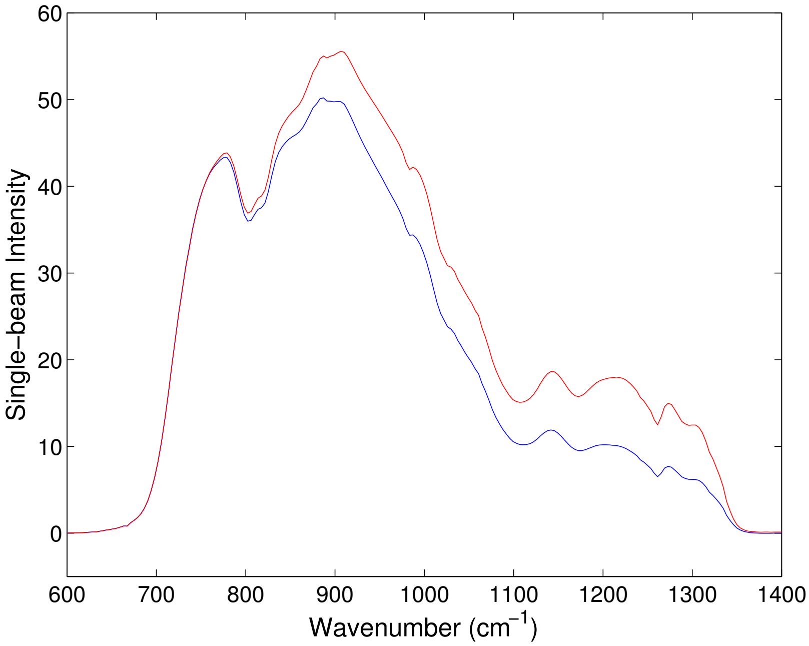
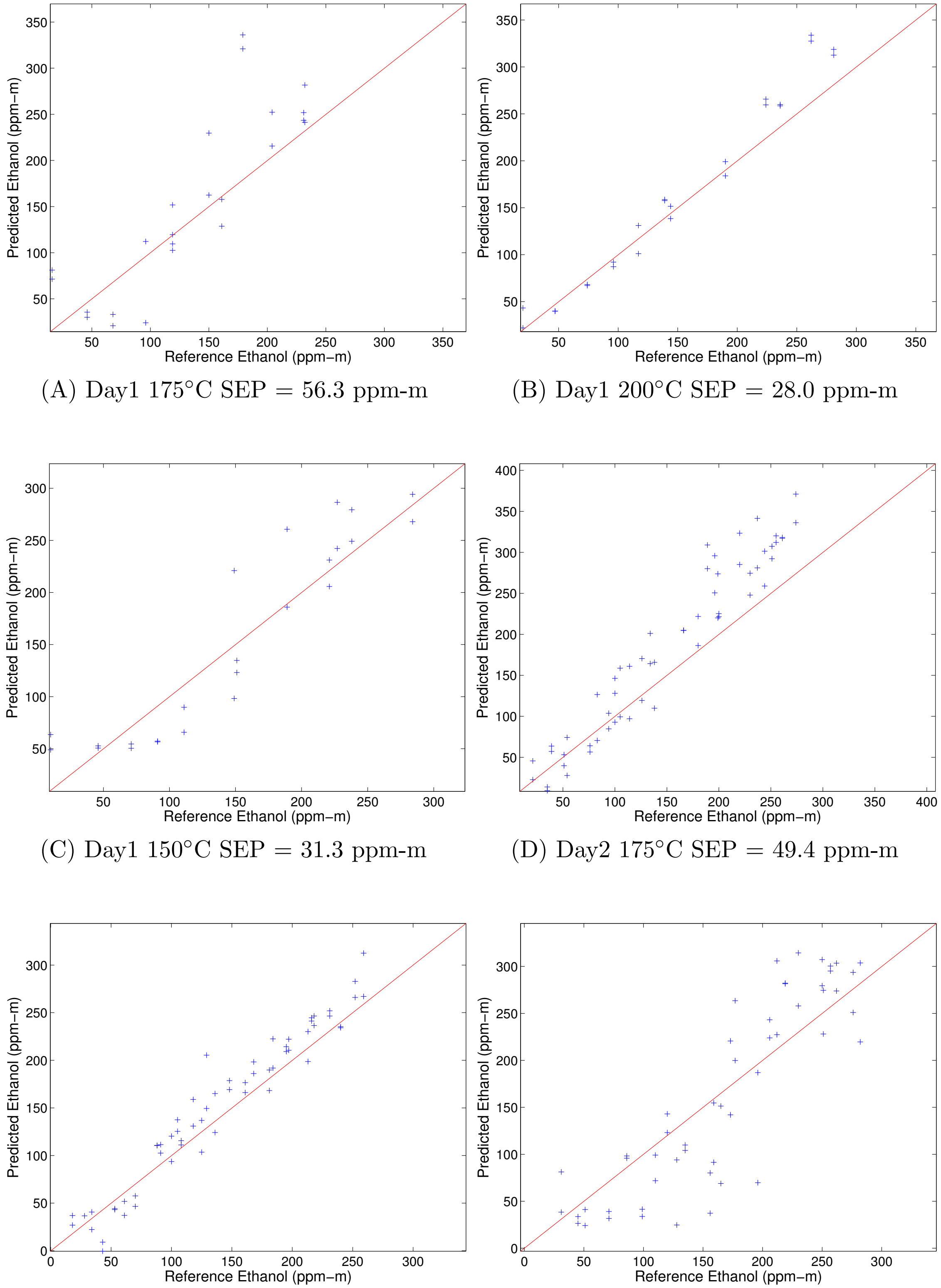


Figure 4.9. Two replicate spectra of a laboratory ethanol sample collected on day 2 (concentration = 1181.9 ppm-m). The spectral intensities between 700 and 800 cm^{-1} are consistent, but the intensities begin to deviate from 800 cm^{-1}. Variation in the single-beam intensities is a reflection of instability in the experimental measurement.



(A) Day1 175°C SEP = 56.3 ppm-m

(B) Day1 200°C SEP = 28.0 ppm-m

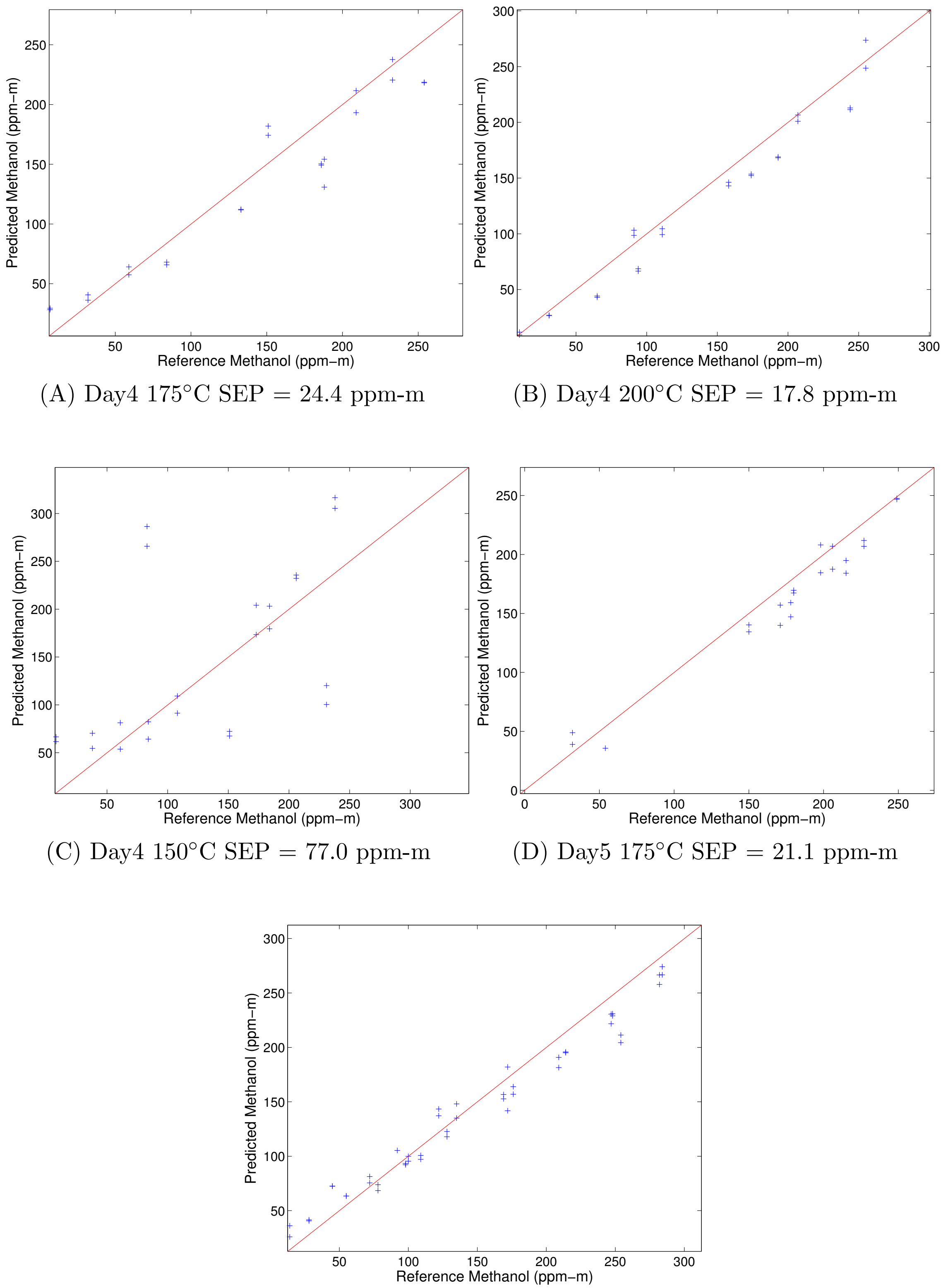
(C) Day1 150°C SEP = 31.3 ppm-m

(D) Day2 175°C SEP = 49.4 ppm-m

(E) Day3 200°C SEP = 23.3 ppm-m

(F) Day3 150°C SEP = 46.5 ppm-m

Figure 4.10. Correlation plots for ethanol predictions with field spectra on each day with different stack temperatures using the PLS model based on synthetic calibration spectra. The spectral range was 1150 – 950 cm^{-1} with two PLS factors.



(A) Day4 175°C SEP = 24.4 ppm-m

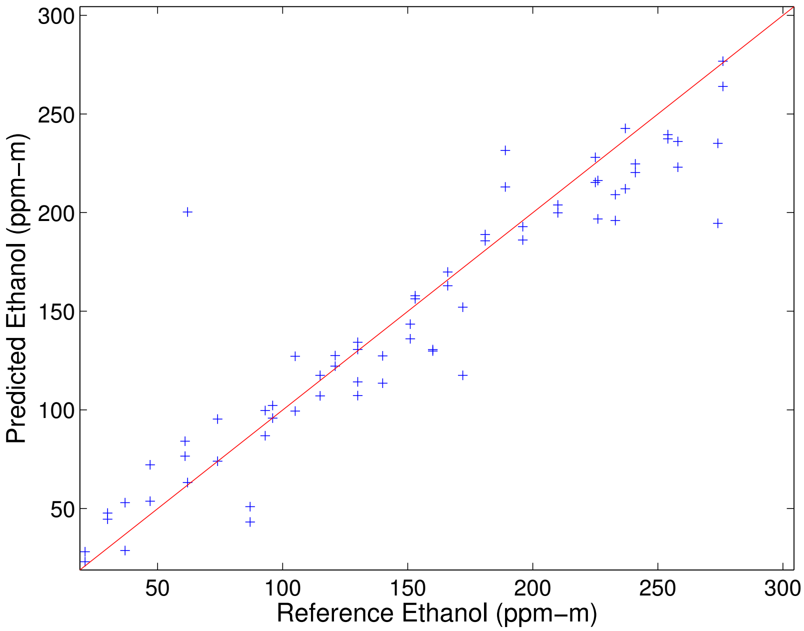
(B) Day4 200°C SEP = 17.8 ppm-m

(C) Day4 150°C SEP = 77.0 ppm-m

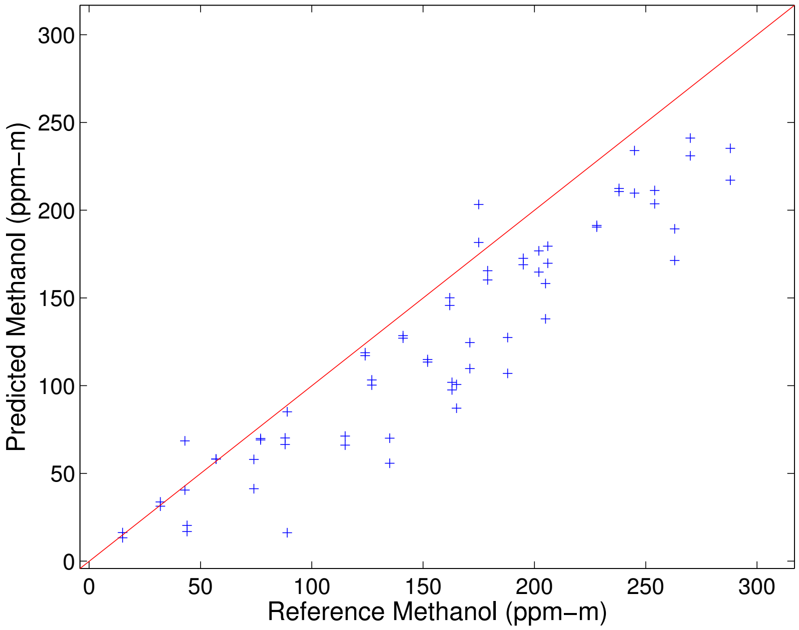
(D) Day5 175°C SEP = 21.1 ppm-m

(E) Day6 175°C SEP = 18.3 ppm-m

Figure 4.11. Correlation plots for methanol predictions in field spectra on each day with different stack temperatures using the PLS model based on synthetic calibration spectra. The spectral range was 1100 – 950 cm^{-1} with two PLS factors.



(A) Day5 175°C SEP = 27.5 ppm-m Ethanol



(B) Day5 175°C SEP = 41.0 ppm-m Methanol

Figure 4.12. Correlation plots for ethanol (A) and methanol (B) predictions for releases of mixture samples on Day 5 at a stack temperature of 175 °C using the PLS model based on synthetic calibration spectra. The spectral range was 1150 – 950 cm^{-1} for ethanol and 1100 – 950 cm^{-1} for methanol. Both models used three PLS factors.

CHAPTER 5
CALIBRATION AND UPDATING STRATEGY BASED ON PARTICLE SWARM OPTIMIZATION OF DIGITAL FILTERING AND PARTIAL LEAST-SQUARES MODEL PARAMETERS: APPLICATION TO CONTINUOUS MONITORING OF GLUCOSE BY NEAR-INFRARED SPECTROSCOPY

5.1 Introduction

With improvements in instrumentation and data analysis techniques, NIR spectroscopy has gained increasing acceptance in analytical chemistry. It is particularly used for quantitative measurement of chemicals with organic functional groups such as C-H, O-H and N-H. Near infrared measurements require little or no sample preparation, can be compatibly used with aqueous samples, and are simple, fast and nondestructive. The technique has found widespread application in pharmaceutical, petroleum, food, agricultural, and clinical analysis.

Significant research efforts have been directed to the development of continuous monitoring based on NIR spectroscopy.[18,91–93] Near infrared spectroscopy is attractive for continuous monitoring applications because of its nondestructive nature and its compatibility with the aqueous sample maxtrixes often encountered in industrial processes.[94–97]

For a continuous monitoring application in an industrial setting, the ruggedness of the instrumentation is of primary importance. While a Fourier transform (FT) spectrometer might be the instrument of choice in a laboratory setting, the presence of moving parts in the interferometer may limit the ruggedness and portability of the instrument for a continuous monitoring application in a non-laboratory setting. In this case, a filter-based instrument is attractive because it can be designed without moving parts and has the potential to be constructed with a smaller footprint. To maintain the required optical performance, however, the filter-based instrument must achieve the same potential spectral S/N ratio and optical throughput that would be obtained with a conventional laboratory spectrometer such as an FT instrument. In this chapter, a filter-based spectrometer constructed with an acousto-optic tunable filter (AOTF) was employed in the spectral collection.

As introduced in Chapter 2, the AOTF is an all-solid state, electronically driven device that uses the acousto-optic effect to diffract propagating light through an anisotropic crystal bonded with acoustic transducers. The incident light is diffracted according to the refractive index from the acoustic wave launched into the crystal. Changes in the acoustic wave alter the diffraction properties of the optical material. Consequently, wavelength selection is made. The acoustic frequency can be changed at electronic speeds, thereby enabling a fast wavelength scan.

The principal drawback of NIR measurements is the occurrence of weak spectral features that are also broad and highly overlapped. Any quantitative calibration must be based on the information from multiple wavelengths, thereby requiring multivariate modeling to be performed. The partial least-squares (PLS) regression method was applied in this chapter to implement quantitative calibration models.

One of the consequences of the requirement for multivariate calibration models is the necessity for estimating several model coefficients. The more coefficients that are required in the model, the greater the likelihood that the model will go out of calibration at some time in the future. This degradation in the calibration can occur because of variations in the instrumental response (i.e., “instrumental drift”), changes in physical parameters associated with the data collection (e.g., temperature), or chemical changes in the sample matrix. The degradation potentially gets more serious with increased time between the collection of the calibration data used to build the model and the subsequent collection of spectra to which the model will be applied.

To address such problems of calibration instability, an obvious solution is to completely recalibrate the model, thereby incorporating any new spectral features that have arisen in the data. However, this approach is time-consuming, expensive and not compatibly used in real-world continuous monitoring applications. A more workable strategy involves simply updating the calibration model or employing data preprocessing methods that serve to remove the effects of any new sources of spectral variation.

Calibration model updating strategies attempt to incorporate new features of the prediction data into the calibration model without performing a complete recalibration.[84] Usually a small set of updating samples is collected to guide the remodeling or optimization of the calibration model. It may be possible simply to acquire spectra of a single blank sample to add to the previous calibration spectra, followed by recalculation and reoptimization of the model.[98]

In terms of methods for data preprocessing, signal processing techniques have been widely studied for use in the removal of spectral variation that could negatively affect calibration and prediction performance.[43,45,59,99] In this chapter, digital filtering methods are examined for use in this application.

If one considers the combined tasks of (1) multivariate modeling, (2) data preprocessing and (3) any necessary model updating, the issue of parameter optimization becomes of paramount importance. Each of the three tasks noted above has a requirement for optimization of parameters inherent to the method. Unless a global optimization strategy is implemented, there is no guarantee that a truly optimal solution has been found. For example, the selection of which spectral wavelength points to include in the multivariate model may be dictated by the characteristics of the preprocessing method chosen.

One optimization approach involves testing all combinations of a selection of key parameters involved in the three modeling tasks outlined above. This "grid search" strategy is only practical if the number of parameters and the number of levels of the parameters needed to be evaluated is small. This method is not realistic for multiple-step calculations, especially when the preprocessing step (e.g., digital filtering) has a large number of parameter combinations. In this research, an alternative method, particle swarm optimization (PSO) is performed to select the optimal parameter set for both a preprocessing digital filter and the multivariate model.

5.2 Experimental

5.2.1 Data Set Design and Sample Preparation

The spectral data used in this study were acquired by our collaborators at ASL Analytical, Inc. Two sets of four-component mixture solutions were prepared through the use of a flow system based on a set of custom computer-controlled peristaltic pumps. The first set of sample solutions contained glucose ($\geq$99 %, Sigma-Aldrich Co., St. Louis, MO), sodium L-lactate (99%, Sigma), glycine ($\geq$98.5%, Sigma) and L-lysine ($\geq$98%, Sigma). This set of data is termed Group 1. The other set (Group 2) of four-component sample solutions consisted of glucose (Sigma), sodium L-lactate (Sigma), urea (Sigma) and creatinine (Sigma).

All samples were prepared in a pH 7.4, 0.1 M phosphate buffer solution. The buffer solution was prepared by dissolving an appropriate amount of NaH_2PO_4 (Sigma) in reagent grade water purified from a water purification system and titrating to pH 7.4 with 50% w/w NaOH (Sigma). The stock solution of each component was prepared by weighing an appropriate amount of each constituent and diluting with buffer. The concentration of the stock solution for each constituent was 60 mM.

All the mixed solutions were obtained by pumping the stock solution of each component at a certain rate and mixing with buffer solutions to get a specific concentration for each constituent. The flow rate through the sample cell was 1.5 mL/min. The rate of each pump was varied to obtain the desired mixture compositions. The concentration of each component in the mixtures was randomly generated and the range was from 1.0 to 30.0 mM. For Group 1, the correlation coefficients computed

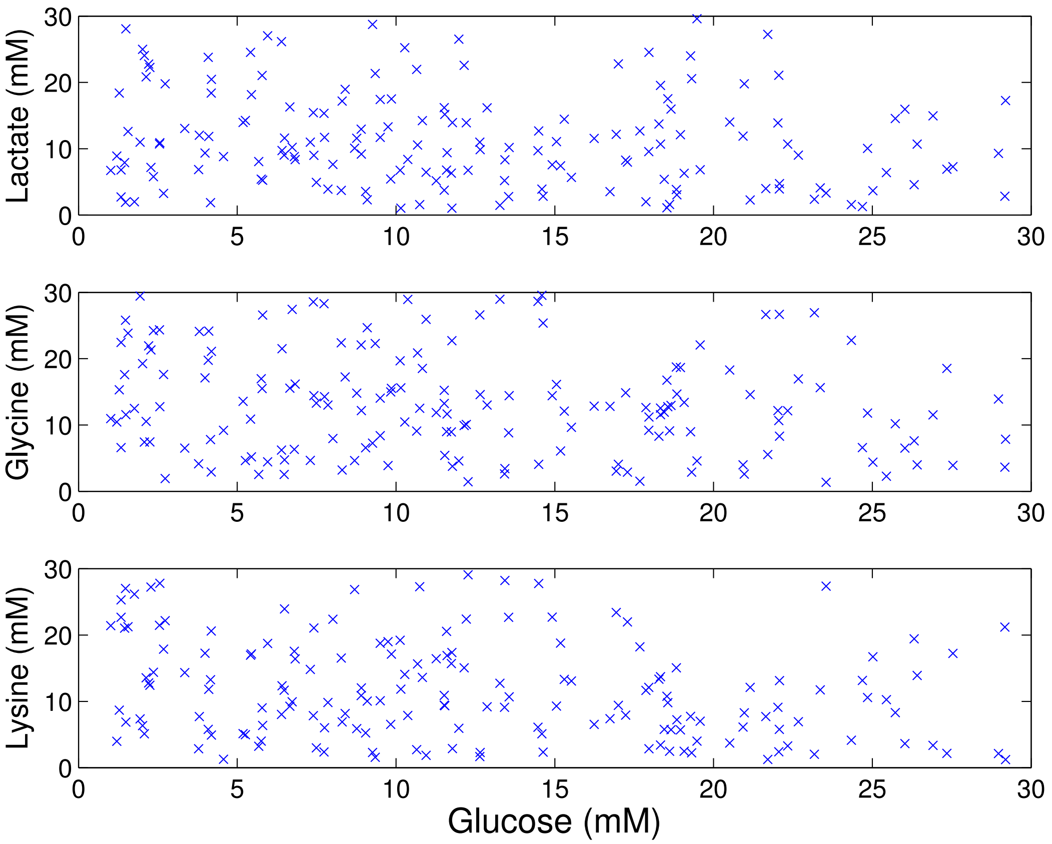


Figure 5.1. Concentration correlation plots between glucose and the other three components for all samples in Group 1.

between pairwise combinations of the concentration profiles of the constituents were all below 0.1. Figure 5.1 displays correlation plots between the glucose concentrations and the concentrations of the other components. The randomly scattered points imply that the correlations between these components were minimal. Analogous results were obtained for Group 2.

5.2.2 Instrumentation and Apparatus

All of the spectra were collected with a custom AOTF-based spectrometer. Figure 5.2 shows the schematic diagram of the instrumental setup. The light source was provided by a 12 W tungsten-halogen lamp. An optical cell composed of Teflon tubing with a path length of 1.3 mm was used for sample flow. In the AOTF, a TeO_2 crystal and a radio frequency (RF) synthesizer (Brimrose model TEAF5-2.0-2.5-EH system, Brimrose Corp. of America, Sparks, MD) with a 5 mm^2 optical aperture were used to provide wavelength tuning from 2000 to 2640 nm (3800 to 5000 cm^{-1}). There was a thermoelectric control unit installed to control the crystal temperature. A beam splitter was applied to divide polarized light passed through the AOTF into sample (90%) and reference (10%) beams. The intensities of the light beams passing through the sample and the open-beam reference channel were detected by two-stage thermal electrically cooled InGaAs detectors (Teledyne Judson Technologies, Montgomeryville, PA) coupled with a low-noise power supply and an integrated pre-amplifier. The detected signals were then transferred to a computer through a 16-bit analog-to-digital interface board (National Instruments Corp., Austin, TX).

A microcontroller (PIC16F877 microchip) running at 20 MHz on a custom circuit board was used to connect the computer with the AOTF for controlling the signal generation. The AOTF was driven by a signal synthesizer capable of a clock rate of 300 MHz (Analog Devices AD9852 DDS) with an operation of chirp mode for spectral collection. The output was then amplified with a 4-W power amplifier (IntraAction model PA-4, IntraAction Corp., Bellwood, IL) operating over a frequency range of 10 – 100 MHz.

5.2.3 Spectral Data Collection and Partitioning

The spectra in Group 1 were collected continuously over two days. Before collecting sample spectra, buffer solution was flowed through the cell and a total of 1551 buffer spectra were measured. To make a sample, each stock solution was pumped at an appropriate rate and the solutions were mixed in a mixing chamber equipped with a magnetic stirrer. The mixed solution was then flowed through the sample cell. For each sample, approximately 10 minutes were taken to collect spectra continuously. Each spectrum was based on a 15-second co-add, corresponding to 600 scans of the AOTF. At the same time, an air reference spectrum was collected for each sample spectrum.

It usually took four minutes for the composition of the sample to stabilize after changing pump speeds. Spectra collected during this time period were discarded and those from the later six minutes were used. According to the scan rate of 15 seconds per spectrum, there were 24 spectra usable for each sample. With the co-addition of every eight consecutive spectra, each sample had three replicate spectra. Buffer spectra, which were collected before the sample collection, were similarly averaged and used as backgrounds in the calculations described below. In total, 168 samples were collected for the short-term study. The first 126 samples (75%) were used to build calibration models, and the last 42 samples were used as prediction data for model validation. Figure 5.3 shows the glucose concentration profile of the samples in Group 1.

The spectra collected in Group 2 were collected over more than one month for

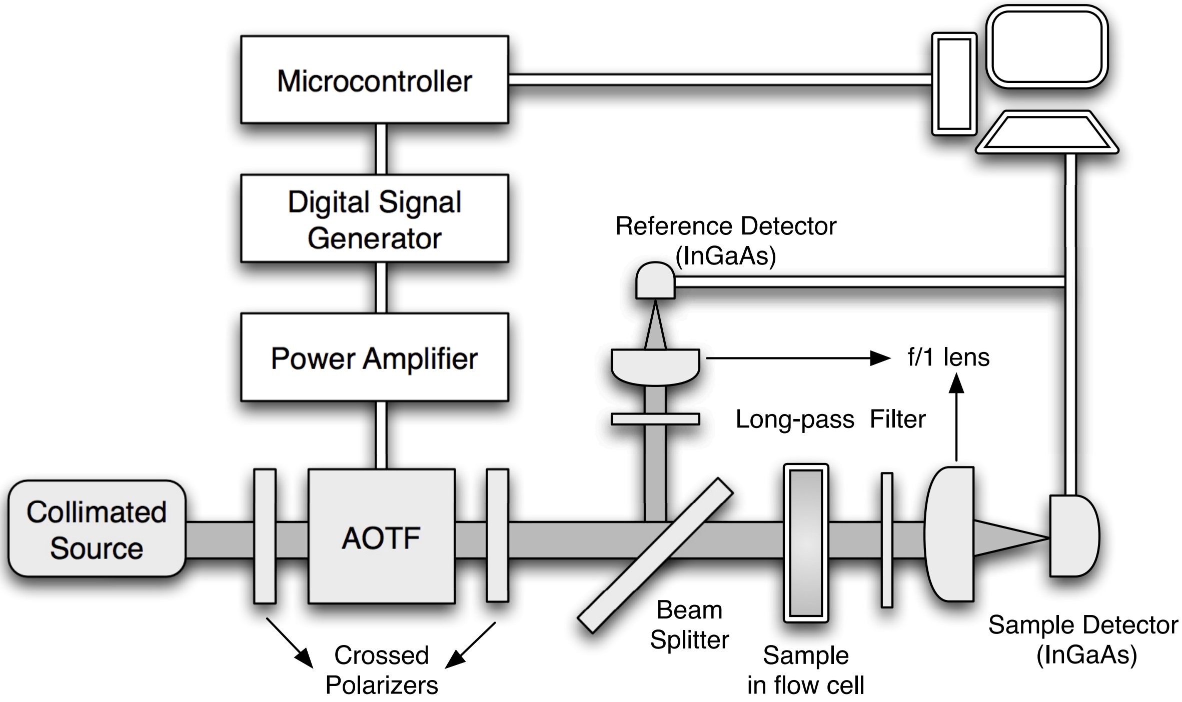


Figure 5.2. Schematic of AOTF-based spectrometer used in collecting spectra.

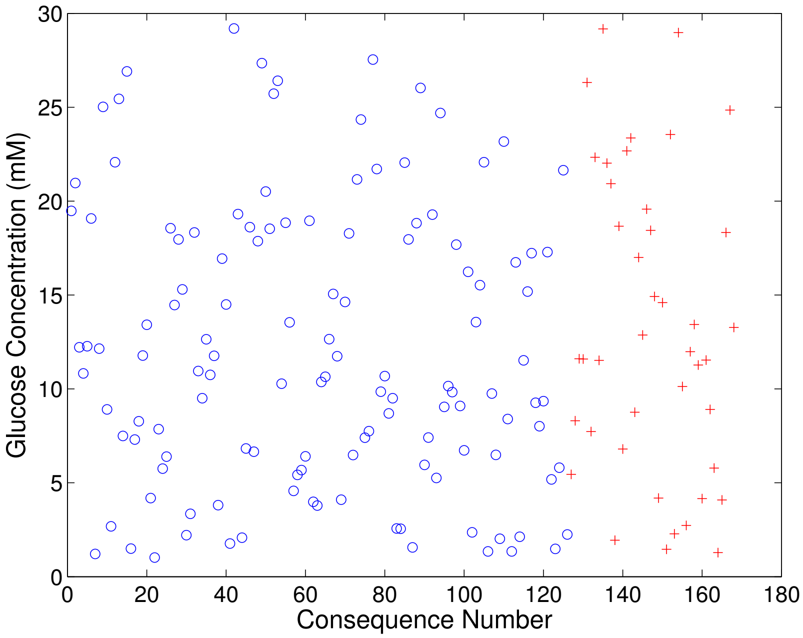


Figure 5.3. Glucose concentration (mM) profile of samples in Group 1. Circles represent the calibration points and crosses are prediction data points.

Table 5.1. Sampling protocol of data collection of samples in Group 2

Data set	Days/Weeks after Calibration	Numbers of Samples	Mixed Sample	Pure Component	Buffer
Calibration	1, 2	168	All mixed samples		
Pred 1	8/1	26	16	8	2
Pred 2	13/2	28	20	4	4
Pred 3	20/3	28	20	4	4
Pred 4	27/4	28	20	4	4
Pred 5	30/4.5	21	18	0	3
Pred 6	50/6	15	12	0	3

the purpose of a stability test. As with the data collection in Group 1, each sample solution was measured for 10 minutes and spectra from the first four minutes were discarded. Every eight spectra were co-added across the last six minutes and three replicate spectra were obtained for each sample. The samples in the calibration set were collected continuously over two days. Prediction data sets were then measured over the following 43 days. The first prediction set was collected seven days after the calibration. There were in total 3458 buffer spectra collected before and after the collection of the calibration samples. In the prediction data sets, some of the samples contained mixtures and some were pure components. Buffer spectra were also counted as samples. Table 5.1 shows the summary of the sampling protocol used in the collection of spectra in Group 2.

The AOTF spectra were collected over 3800 to 5300 cm^{-1} with a point spacing of 0.57 cm^{-1}. Each point in the spectrum was the mid-point of a square-shaped spectral band-pass filter, which was approximately 24 cm^{-1} in width. One single-beam spectrum was obtained in 15 ms by tuning the center position of the filter rapidly.

5.2.4 Calculations

All computational work was performed under MATLAB (version 7.4, The MathWorks, Inc., Natick, MA) running on a Dell Precision 670 workstation (Dell Computer Corp., Austin, TX) operating under Red Hat Enterprise Linux WS (Red Hat, Inc., Raleigh, NC). The digital filtering and PSO calculations were implemented with the Matlab signal processing toolbox (version 6.7, The MathWorks, Inc.) and the public-domain Particle Swarm Optimization Toolbox developed by Brian Birge (version 2.5) and available through the file exchange maintained by The MathWorks, Inc.

5.3 Results and Discussion

5.3.1 Data Characterization

In this experiment, for both groups of data, glucose was the analyte of interest. There are three combination bands present in the 5000 – 4000 cm^{-1} region. The broad band around 4700 cm^{-1} corresponds to an O-H combination absorption. The other narrower bands at 4400 cm^{-1} and 4300 cm^{-1} are from C-H stretch-bend combinations. The glucose absorption bands are much weaker than those of the other components and highly overlapped. Therefore, the qualification and quantification of glucose in such an environment becomes challenging.

The noise level among replicate spectra can be used to evaluate the quality of the collected data. With each group of three sample single-beam spectra, three noise spectra in absorbance units (A) were computed by taking the ratios of each possible combination of spectra (1 vs. 2, 1 vs. 3, 2 vs. 3):

$$A_i = -\log \frac{P_{rep1,i}}{P_{rep2,i}} \tag{5.1}$$

In Eq. 5.1, $P_{rep,1}$ and $P_{rep,2}$ denote the intensity at point i of the single-beam spectra from a given pair of replicates. The resulting absorbance spectrum is termed a '100% line'. In theory, the 100% transmittance profiles of each two replicate pair should have no instrumental noise. However, due to the instrumental variation or temperature changes occurring during the data collection, the 100% lines always have fluctuations and the curve can be modeled by a polynomial function.

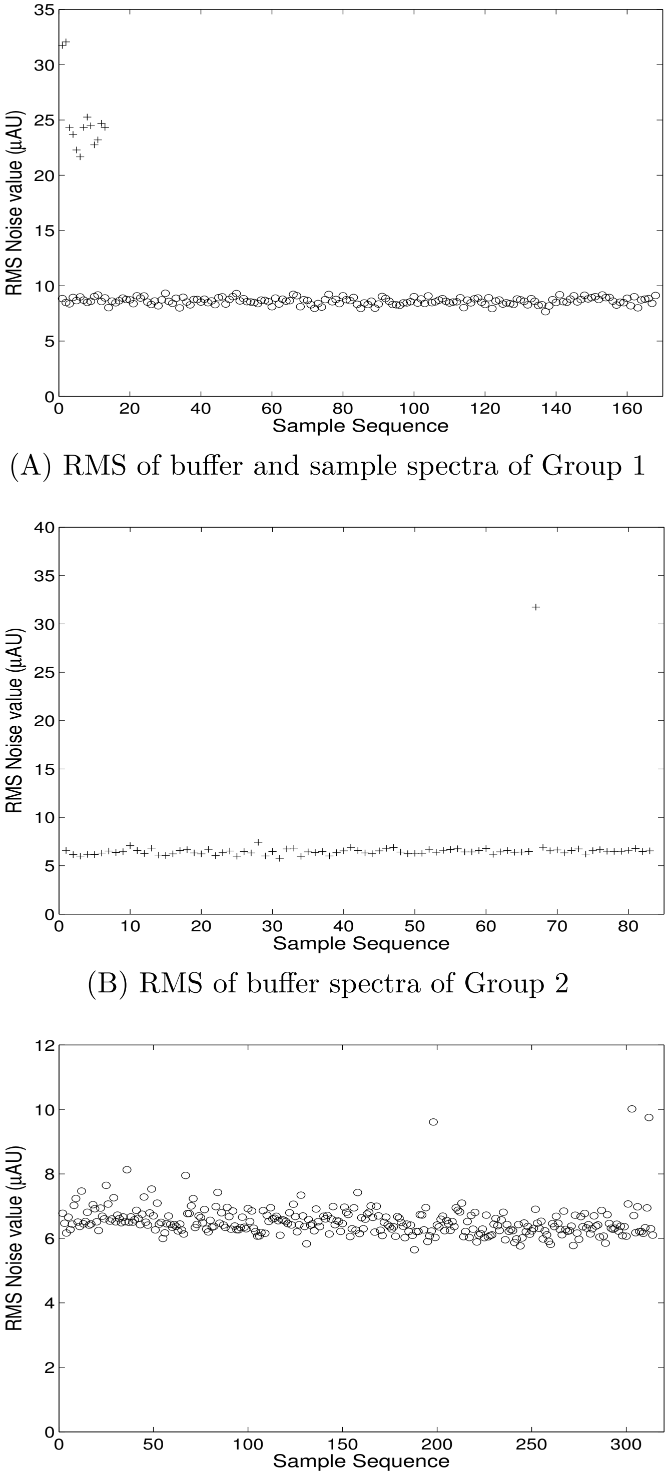
The noise value can be estimated by the root-mean-squared (RMS) error of the deviations between a fitted quadratic function and the computed 100% lines of the three replicate noise spectra. Eq. 5.2 shows the calculation of RMS error, where d_k represents the deviation among the spectrum at the k^{th} spectral frequency, n is the number of wavelengths (or wavenumbers) involved in the calculation, df represents the loss of degrees of freedom in the polynomial fitting (df = 2 for a quadratic fit), and "+ 1" indicates the loss of one degree of freedom corresponding to the use of an intercept term in the model.

$$\text{RMS} = \sqrt{\frac{\sum_{k=1}^{n} d_k^2}{n - (df + 1)}} \tag{5.2}$$

In this study, the spectral range from 4500 to 4300 cm^{-1} was used for the quadratic fitting and RMS error calculation. As introduced previously, this range covers the important C-H combination band of glucose at 4400 cm^{-1}, which can be used for glucose qualification and quantification.[100] All of the estimated errors were computed in micro-absorbance units (μAU).

For each calibration and prediction data set, the RMS error values were estimated for each sample with the three replicates. The pooled noise value was calculated for each data set by taking the mean of the RMS noise values for all samples. The RMS noise of the buffer spectra collected 10 minutes before sample collection was also calculated. Panel A in Figure 5.4 plots the RMS noise values of both sample and buffer spectra of Groups 1 and 2. The crosses represent the noise of the buffer samples, and the circles are the noise values of the sample spectra. The mean RMS noise of the sample spectra was around 10 μAU, and that of the buffer spectra was around 25 μAU. In collecting the data in Group 2, the light source was changed. There was a new bulb holder installed and the signal was maximized by adjusting the alignment. Therefore, from panels B and C in Figure 5.4, the RMS noise of Group 2 improved because of an increase in the light throughput. The noise values of both buffer and sample spectra were around 6 μAU.

In the flowing system used in the data collection, solutions were pumped from the stock solution into the sample cell. The occurrence of air bubbles in the flow cell is common and required the use of a detection algorithm to identify and remove spectra altered by the presence of bubbles. Figure 5.5 depicts the comparison of a single-beam spectrum with an air bubble in the sample cell and a normal single-beam



(A) RMS of buffer and sample spectra of Group 1

(B) RMS of buffer spectra of Group 2

(C) RMS of sample spectra of Group 2

Figure 5.4. RMS noise values of sample and buffer spectra. A. Buffer and sample RMS noise values of data in Group 1 (cross - buffer, circles - sample). B. Buffer RMS noise values for data in Group 2. C. Sample RMS noise values for data in Group 2.

spectrum collected without the presence of air bubbles. Comparison of the spectra shows clearly that the single-beam intensities are altered in the regions below 4200 cm^{-1} and above 4800 cm^{-1} when air bubbles are present. This is caused by the larger light throughput in these regions for air than for water.

To identify spectra contaminated by air bubbles, the ratio of the single-beam intensities at 4570 cm^{-1} and 4100 cm^{-1} was calculated for each spectrum. This ratio decreases when air bubbles are present.

Figure 5.6 plots the ratio values of all spectra collected in Group 1. For most of the spectra, the values were above 65. For 33 spectra, the values were lower than 50. These abnormal spectra were removed with a threshold value at 60 before the calculation of RMS noise values and co-addition of spectra for the construction of replicates. The same procedures were taken for data collected in Group 2.

5.3.2 Data Analysis Strategy

In this chapter, a robust and stable calibration modeling method was implemented. Due to the weak spectral bands and overlapped features, signal processing (e.g., digital filtering) was employed to help isolate useful analyte information. With the combination of digital filter design and calibration model development, parameter optimization is crucial. A population based optimization method was investigated.

The raw data were collected at a point spacing of 0.6 cm^{-1}. This corresponds to ~2000 points per spectrum over the full spectral range. Taking all of the points in the data analysis is unnecessary and also increases the calculation load. In order to reduce time in calculation, all spectra were deresolved to around 4 cm^{-1} point spacing. In the data collection, part of the incident light was directed to the air reference after passing through the AOTF. Therefore, for each sample spectrum, an air spectrum was obtained. The air reference coupled with each sample spectrum can help monitor instrumental drift during the data collection. This data collection strategy is potentially valuable for maintaining spectral stability during long-term monitoring. There were also buffer spectra collected before and after the collection of the sample spectra. In the data analysis, both buffer and air spectra were taken as the background in the evaluation of potential calibration models.

Digital filtering was applied in the data preprocessing procedure. By Fourier analysis, a spectrum can be modeled as the sum of sine and cosine functions across a finite bandwidth. Given this frequency-dependent model, a bandpass filter can be used to suppress signals corresponding to component frequencies that are not associated with useful spectral information.

For example, random noise changes rapidly and thus consists of a series of narrow signals. The Fourier analysis of such signals requires frequencies across the total bandwidth of the data. Spectral variation associated with wide features (e.g., baseline drift) requires fewer frequencies to model, and these are concentrated in the lower-frequency portion of the bandwidth. Thus, by applying a bandpass filter to the data which passes only intermediate frequencies, much of the spectral information associated with both very wide and very narrow features can be removed. This has the dual benefit of helping to suppress both baseline variation and random noise.

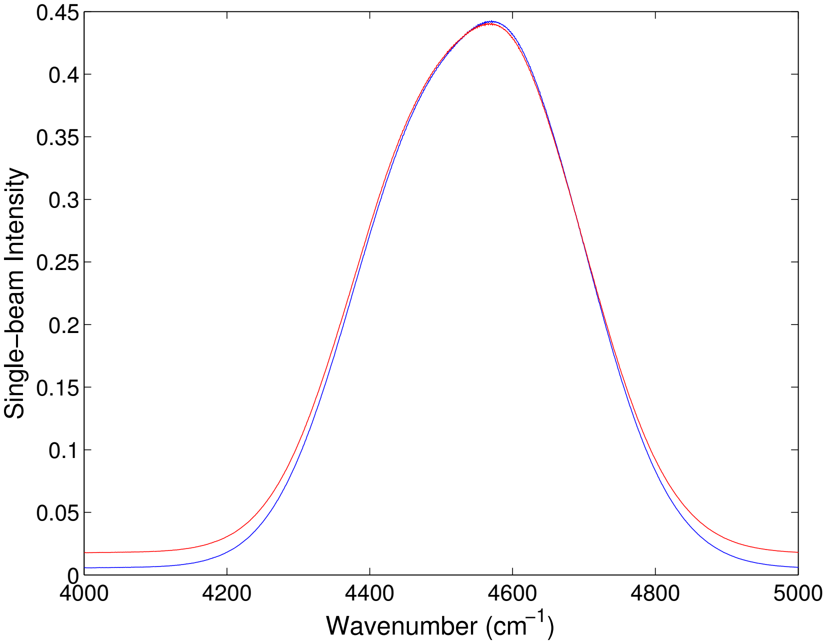


Figure 5.5. Comparison of single-beam spectra of samples with air-bubbles (red) and those without air-bubbles (blue).

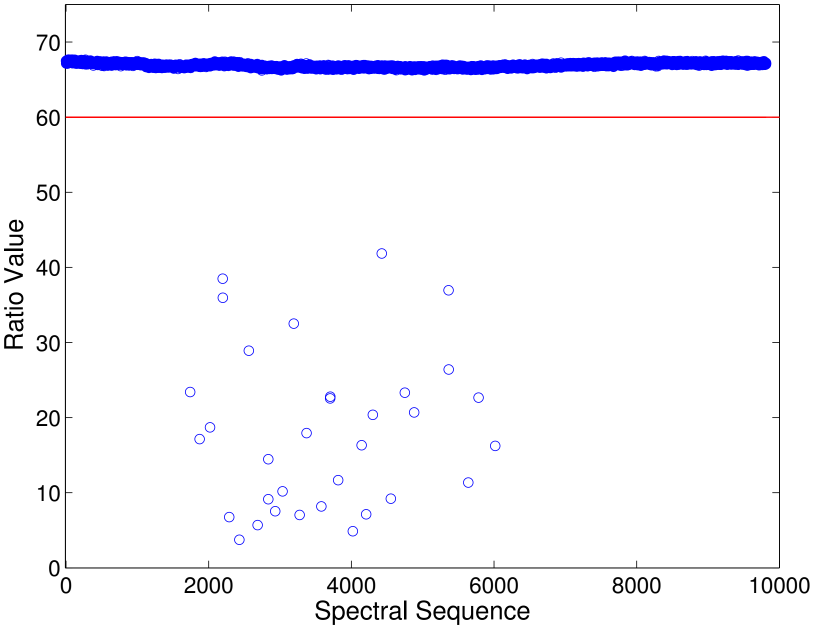


Figure 5.6. The ratio of intensities at 4570 cm^{-1} and 4100 cm^{-1} of all original single-beam spectra in Group 1. The red line shows the threshold value of 60 used to identify spectra altered by the presence of air bubbles.

One of the classical infinite impulse response (IIR) filters, the Chebyshev Type II bandpass filter, was investigated in this study. Compared to finite impulse response (FIR) filters, IIR filters can achieve an excellent approximation to the desired frequency response with a much lower filter order, thereby requiring the use of fewer spectral points in the application of the filter. This advantage arises because the IIR filter uses both unfiltered data points, as well as previously filtered points, in the estimation of the filtered intensity of a given data point. The Chebyshev type II filter design was chosen because it provides a fast roll-off between the passband and stop-band, consequently leading to precise and flexible control of the frequency output of the filter.

The filtered spectra were used as input to the PLS algorithm. The number of latent variables (factors) is the principal control parameter associated with the PLS method, specifying the number of levels of decomposition of the input spectral data matrix. The PLS method usually gives a better calibration performance if more factors are added. This is because additional added variables will always be able to explain more variation in the calibration data that is being used to derive the model. However, the variance explained by the additional factors could be noise or other irrelevant information that may not necessarily be present in the prediction data to which the developed model will be applied. This problem is termed over fitting. If those factors are included in the model, since the model has been skewed to the calibration data, the prediction could be biased or ruined. Therefore, to avoid over fitting, an F-test or other statistical test is performed to assess the significance of adding new factors to the model.

In order to select the optimum calibration model, an internal validation method was used in this study. The whole calibration data set was divided into calibration and monitoring sets. From the entire calibration set, 75% was randomly selected as the calibration subset used to build the model, and the remaining 25% was taken as a monitoring set for model validation and optimization. The subsets were selected dynamically to avoid the dependence on how the monitoring data were selected. In optimization, three random drawings of the subsets were tested and the pooled error values were used in calculation of the fitness scores.

5.3.3 Implementation of Particle Swarm Optimization

In optimization, with the combination of a Chebyshev type II band-pass filter and PLS regression, there are a huge number of parameter combinations applicable to the model. Because of the tedious process and large scale of the calculation, a conventional grid search, which requires evaluating all possible parameter combinations to find the optimum solution, is not practical. An iterative numerical optimization method is therefore necessary. In Chapter 3, the PSO algorithm was introduced. Next, how the PSO was implemented with the data analysis strategy will be described.

The first step of PSO is generating the initial population (swarm). Each particle in the population is defined by two vectors, position $\vec{x_i}$ and velocity $\vec{v_i}$. The dimensionality of the particles depends on the number of parameters being optimized.

In the filter design, the Chebyshev type II bandpass filter requires four parameters: (1) the filter order, (2) the desired stopband attenuation, (3) the high-frequency cutoff for the filter bandpass, and (4) the corresponding low-frequency cutoff. For the PLS regression step, the spectral range (i.e., the starting and ending points of a contiguous spectral range) should be optimized. Taken together, this defines a six-dimensional particle.

Because the numerical range of the parameters varies significantly, an integer mapping was used to represent all variables. For example, the spectral range was studied from 4900 cm^{-1} to 4200 cm^{-1}, but the magnitude of the filter order was from 1 to 6. In optimization, for the range from 4900 cm^{-1} to 4200 cm^{-1}, if the decrement of the spectral wavenumber is 10 cm^{-1}, there are 71 possible values in total. To make the integer map, all of the values were assigned to integers from 1 to 71 instead of the wavenumber values themselves. The stopband attenuation was studied from 20 to 100 dB at a step of 5 dB. The frequency cutoff was normalized between 0 and 1, and was investigated from 0.01 to 0.99 with an increment of 0.01.

PSO moves to the evaluation step after the initial population is defined. In this work, 50 particles comprised the population. The purpose of the evaluation step is to locate the particles in the search space in terms of their suitability as solutions to the optimization problem. Each particle is evaluated by the fitness function and a fitness score is obtained.

The fitness function is the key to the optimization. It determines the validity of the ranking of the possible solutions to the optimization problem. The fitness function used in this study is shown in Eq. 5.3.

$$R = \mathrm{SEC} + \mathrm{SEM} + 2|\mathrm{SEC} - \mathrm{SEM}| \tag{5.3}$$

In this equation, SEC is the standard error in concentration of the calibration data and SEM represents the corresponding error for the monitoring set. In computing the value of R that represents the fitness of the particle, the number of latent variables used in the PLS model was also evaluated. When a given particle was tested (i.e., a set of parameter values is evaluated), PLS models were built with all latent variables over the range of 6 to 12. Each model produced a value of SEC and SEM. An F-test at the 95% confidence level was then performed to identify the smallest number of latent variables that produced a value of SEM that was not statistically different from the smallest SEM. This value of SEM along with the corresponding value of SEC was used to compute R.

The last term in the fitness function evaluates the similarity of the calibration and prediction results. If the calibration and monitoring errors (i.e., SEM and SEC) are inconsistent, a penalty is applied to the fitness score. This helps to exclude the case in which the model is skewed toward the calibration results, thereby sacrificing performance in prediction. Particles with lower fitness scores are considered to be better solutions.

After the evaluation of the initial population, the current location of each particle is assigned as *pbest* and the fitness score is recorded. All fitness scores obtained from the population are compared. The position with the best score in the entire swarm is stored as *gbest*. In PSO, the *pbest* and *gbest* are important because they guide all particles moving through the search space.

The third step of PSO is to update the positions and velocities of all particles. All particles move to their new positions according to the changed velocities in the search space. The equations and diagram of the updating procedure have been introduced in Chapter 3. For one particle, a new velocity vector is calculated based on its current position, the distance and direction to its *pbest* and the *gbest*, respectively. The new position is determined by the current position and the new velocity. In this study, the control parameters, a, b_1, and b_2 were selected according to Trelea's study.[101] The inertial constant, a, was 0.6, and the values of the attraction coefficients, b_1 and b_2, were both 1.7.

After all particles are updated, PSO moves back to reevaluate the particles at their new positions (i.e., the updated solutions to the optimization problem). A new set of fitness scores is obtained. For each particle, the current fitness score is compared with its *pbest* and an updated *pbest* is recorded. All fitness scores are compared together with the previous *gbest* to obtain the current *gbest*.

The evaluation and updating steps are repeated for a certain amount of iterations or until the optimal solution (i.e., a specified target fitness score) has been found. In this study, because the monitoring data set was dynamic, a targeted optimum fitness score could not be defined. Initial experiments suggested that little improvement in the fitness score was obtained after 80 iterations. Accordingly, 100 was set as a safe value for the maximum number of iterations.

The occurrence of local optima is a general problem in numerical optimization, in which the solution is only optimal within the local neighborhood instead of within the entire search space. To address this issue, ten different initial populations were generated to obtain 10 optimal solutions. From the 10 solutions, the best set of parameters was selected to build the final calibration model. Prediction was then performed with the parameters corresponding to the best calibration.

5.3.4 Optimization Results for Short-Term Data

The short-term data were collected on two consecutive days. A total of 168 samples were obtained, and divided into calibration and prediction sets chronologically. The concentration profile has been shown in Figure 5.3. Both buffer and air spectra were tested as backgrounds. The buffer background spectrum was the mean spectrum from the last 10 minutes of collection before the first sample spectrum was collected. As shown in Figure 5.4, the first several spectra at the beginning of the 10 minutes had relatively high RMS noise values. These spectra were discarded in calculating the mean spectrum. An air background spectrum was available for each sample spectrum. Air spectra potentially contain more information about the status of the instrumental drift than buffer spectra because they are matched to the individual sample spectra.

Figure 5.7 shows single-beam spectra of sample, buffer and air. In the upper panel, the red line represents the sample spectrum and the blue line is the buffer spectrum. The shapes of the two spectra are similar. The intensity difference is

Table 5.2. Summary of prediction results with buffer and air as background

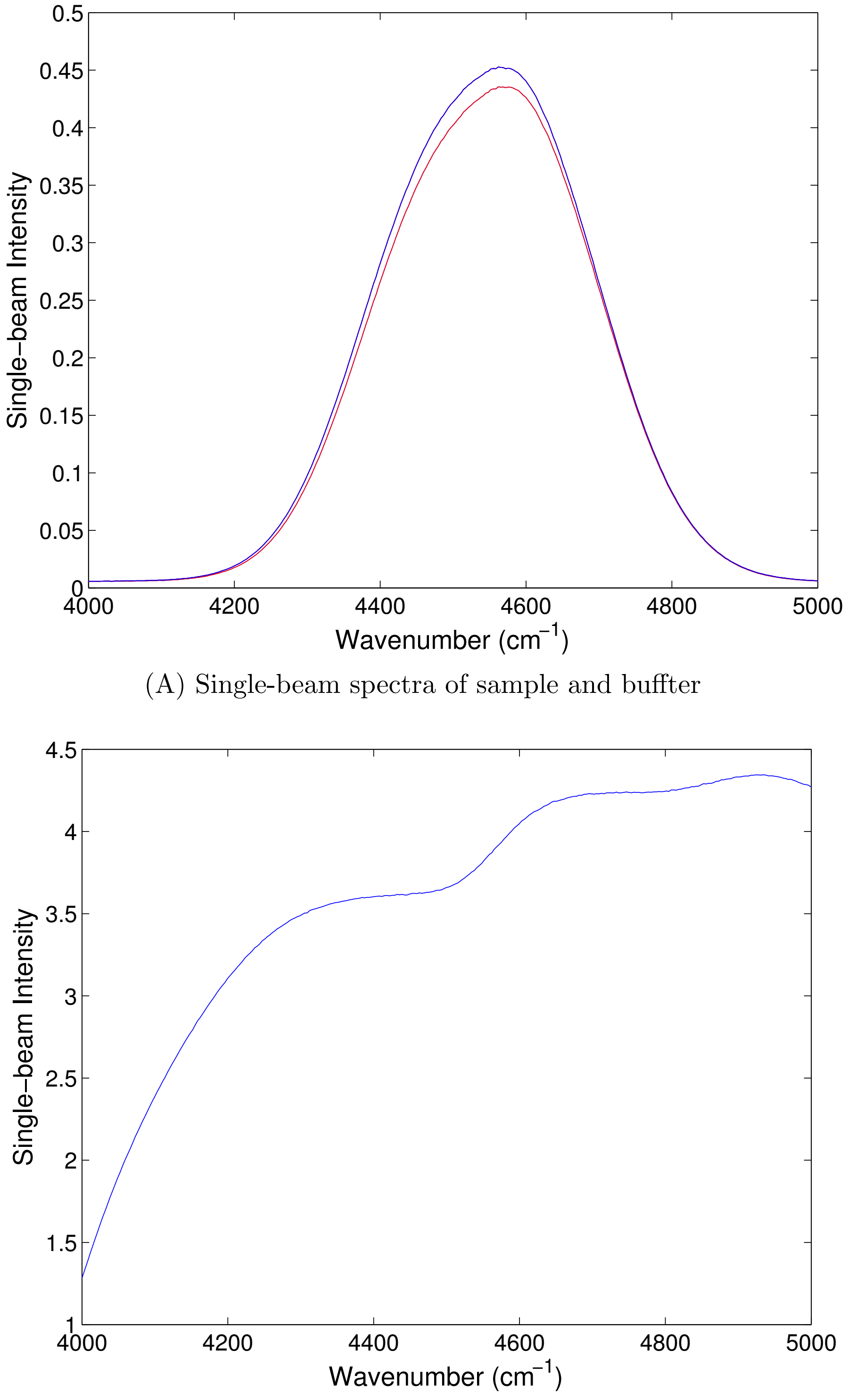
Background	Spectral Range (cm^{-1})	Number of LV[a]	SEC (mM)	SEP (mM)
Buffer	4820 - 4220	10	0.184	0.190
Air	4870 - 4240	10	0.159	0.174

[a] Number of latent variables used in PLS models.

caused by the difference in transmission arising from the differing compositions of the buffer and sample solutions. The lower panel plots the air spectrum, whose shape is different from that of solution. In Figure 5.8, absorbance spectra of a sample are plotted for the case in which buffer (upper) and air (lower) are used as the background. When air is used as a background, the only recognizable features are the tails of the large absorption bands of water centered near 3800 and 5200 cm^{-1}. When buffer is employed as the background, the water absorbance bands are largely subtracted, uncovering the absorption features of the solutes.

Table 5.2 summarizes the calibration and prediction results with spectra using buffer and air as backgrounds. Both backgrounds provide stable calibration and prediction. As discussed previously, however, air spectra carry more information about the instrumental variation. Therefore, both the calibration and prediction errors with air as background are slightly lower, compared to those obtained with buffer spectra.

Figure 5.9 shows correlation plots between the reference glucose concentrations and the predicted values for the calibration (A, C) and prediction (B, D) data sets when buffer (A, B) and air (C, D) were the backgrounds. The high degree of



(A) Single-beam spectra of sample and buffter

(B) Single-beam spectra of air

Figure 5.7. Single-beam spectra. (A) Single-beam spectra of sample (in red) and buffer (in blue). (B) Air single-beam spectrum collected by the AOTF-spectrometer

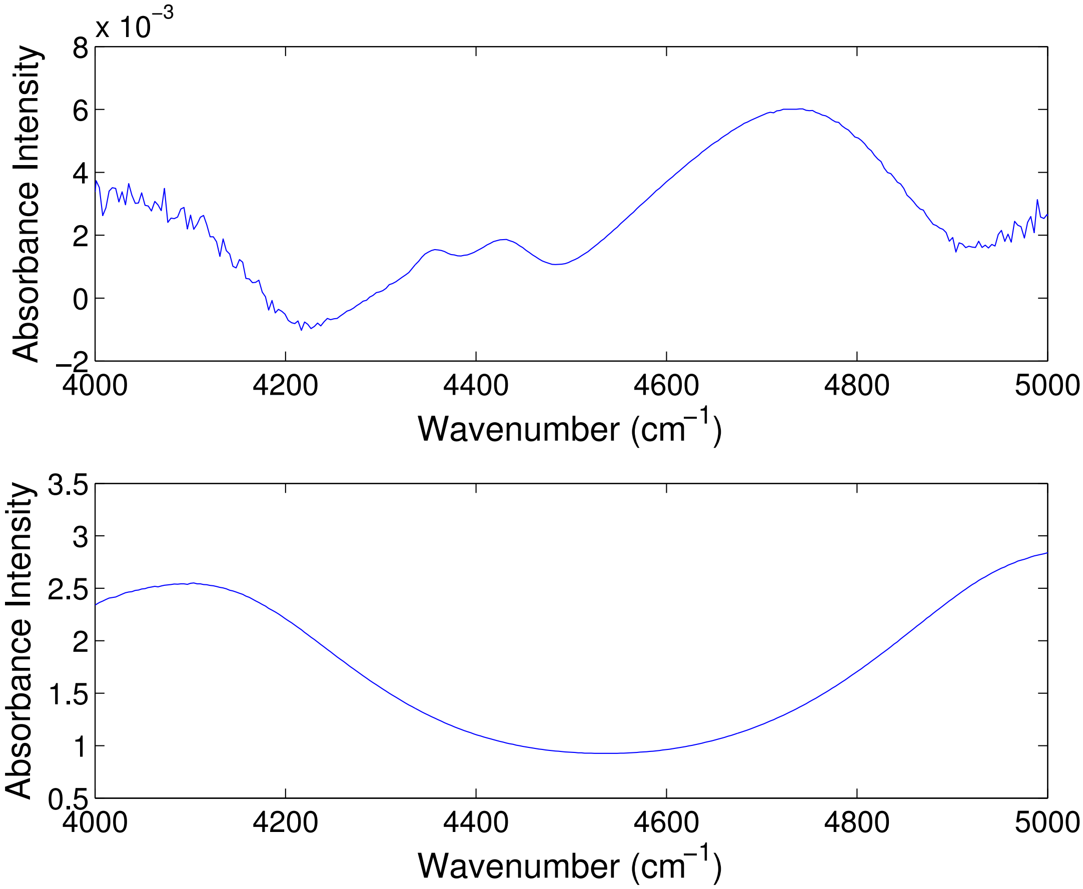


Figure 5.8. Absorbance spectra of a sample when using buffer (upper panel) and air (lower panel) as the background

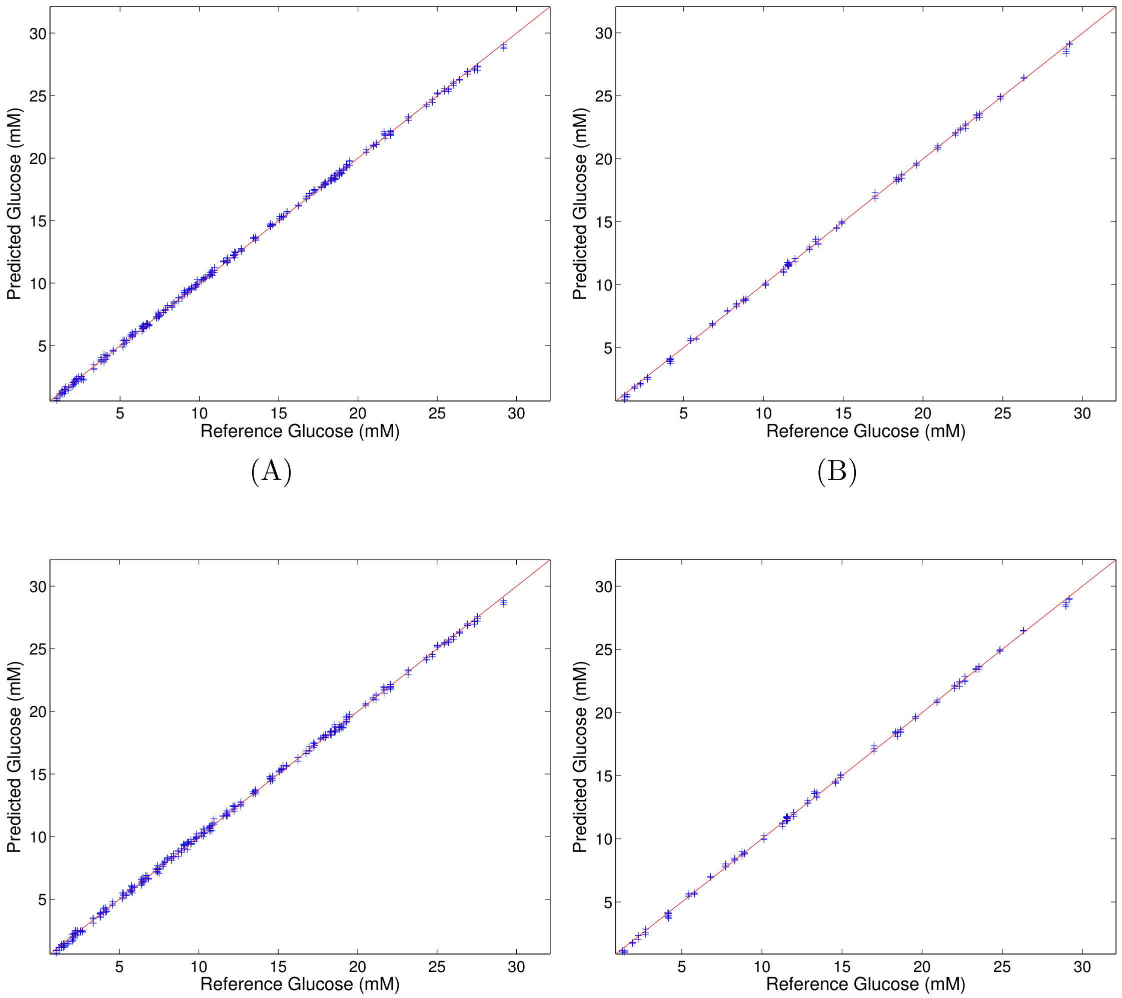
correlation between the reference and predicted values in all four figures indicates good prediction performance of this method when applied to short-term data.

Figure 5.10 displays the frequency responses of the digital filters applied to buffer and air absorbance spectra and example spectra after application of each filter. After digital filtering, part of the spectral information has been removed. The remaining spectra are potentially more related to the analyte feature. The procedure of finding the appropriate frequency cutoffs and other parameters is highly dependent on the fitness function employed with the PSO. In this study, both the calibration and monitoring errors were taken into consideration. If another fitness function were chosen, the optimization results could be different.

5.3.5 Optimization Results for Long-Term Data

Results from the short-term study indicate that this method could extract the specific analyte information from the spectra. However, data in Group 1 were collected only over two consecutive days. The way the data were divided, the prediction set was collected immediately after the calibration data. Although the prediction set was collected after the calibration set, the stability and robustness of this method are not well proved.

For the purpose of continuous monitoring of the analyte of interest, usually a long-term sampling protocol is needed. As mentioned in section 5.1, data in Group 2 were collected over two months for the model stability study. The calibration set was first collected over two days, followed by six prediction sets collected in the following



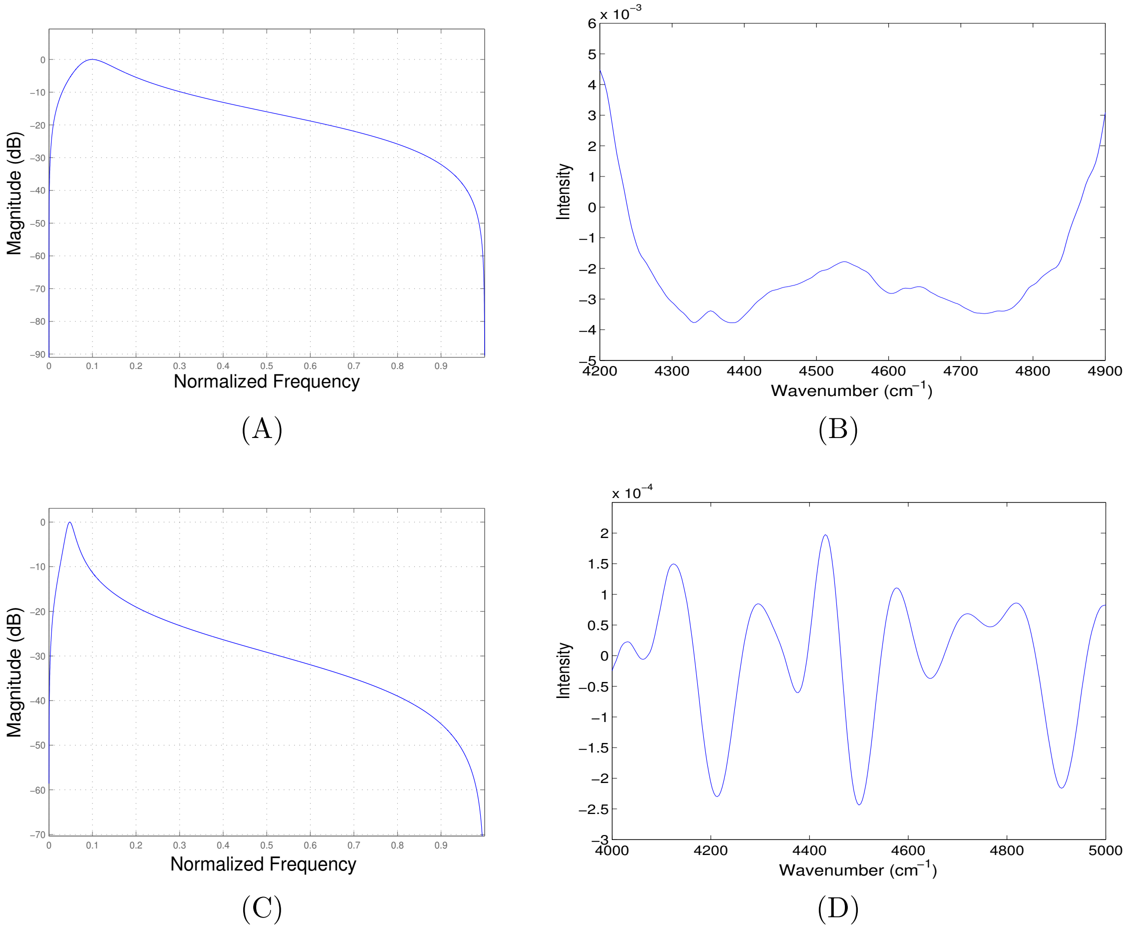
(A)

(B)

(C)

(D)

Figure 5.9. Correlation plots of predicted glucose concentrations versus reference concentrations obtained with the calibration model built with buffer and air absorbance spectra. (A) Calibration correlation of buffer absorbance spectra. (B) Prediction correlation of buffer absorbance spectra. (C) Calibration correlation of air absorbance spectra. (D) Prediction correlation of air absorbance spectra.



(A)

(B)

(C)

(D)

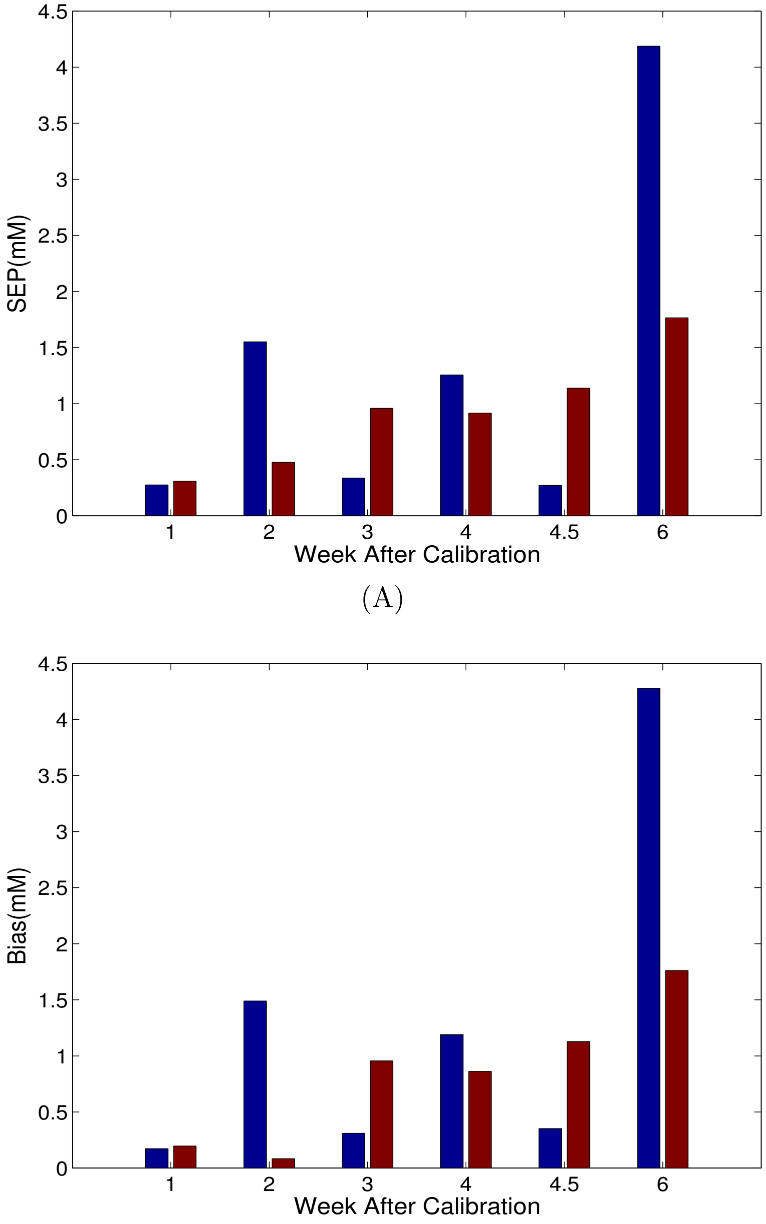
Figure 5.10. Frequency responses of filters applied to buffer (A) and air (C) absorbance spectra and their corresponding filtered spectra for sample spectra computed with buffer (B) and air (D) as backgrounds.

two months. The data collection and partitioning has been discussed earlier in section 5.1.3.

From the comparison results of buffer and air as backgrounds in the short-term study, air absorbance spectra showed a better performance but the differences were small. This comparison was also applied to the long-term data. Figure 5.11 depicts the SEP and bias values obtained when air or buffer was taken as the background. The bias value is the average difference between the predicted and reference concentrations. It shows clearly that, from a long-term point of view, although the buffer spectrum could provide a stable background in some cases, it is highly variable. On the other hand, the air spectrum, collected simultaneously with the sample spectrum, seems to provide a more stable background. This result is expected, since air spectra can monitor the instrumental variation in real time during the data collection. Drift in prediction is still noted over time, however, indicating that the PLS model gradually goes out of calibration, producing an increasingly negative bias in the predicted concentration.

5.3.6 Principal Component Analysis

Principal component analysis (PCA) was performed on the mean-centered air-absorbance spectra before and after digital filtering. A principal component (PC) score plot allows a convenient depiction of the distribution of the spectra across the time span of the data collection. Figure 5.12 shows the comparison score plots. In the unfiltered data, clear clusters of calibration and prediction data sets can be



(A)

(B)

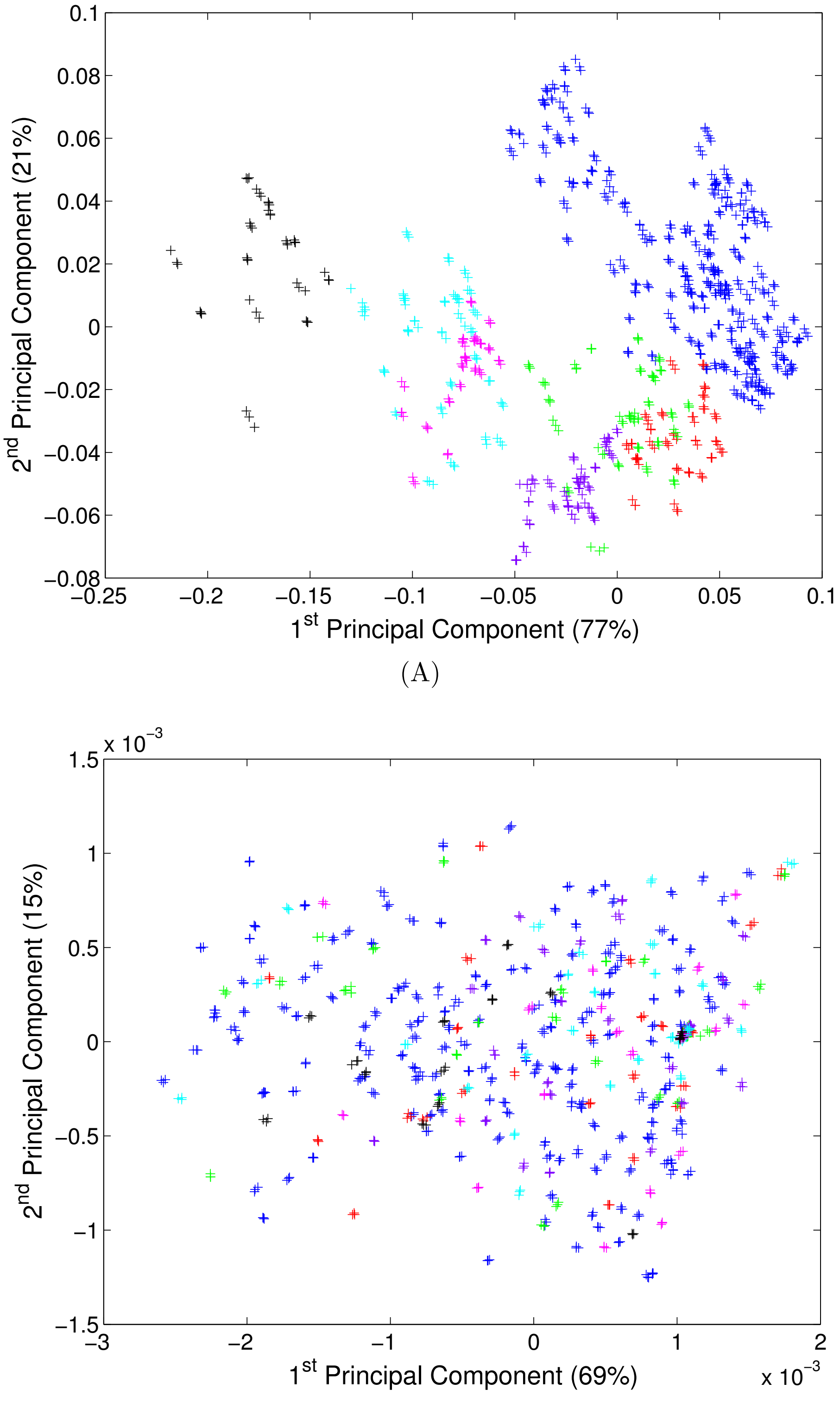
Figure 5.11. Compared SEP and bias plot of the prediction results with buffer or air as the background. (A) shows the SEP values. (B) is the absolute bias value. Blue bar: using buffer absorbance spectra; Red bar: using air absorbance spectra.

observed. The prediction sets deviate from the calibration data over time, which indicates that the information embodied in the calibration set cannot adequately represent the prediction data. Higher SEP values are expected under these conditions. After digital filtering, however, the calibration and prediction data are more scattered, and blended much better, which indicates the calibration spectra are more similar to the prediction spectra. The spectral variation not related to the analyte information has been removed by digital filtering.

5.3.7 Calibration Updating Strategy

As noted in the discussion of Figure 5.11, degradation in prediction performance is observed over time, even with the use of digital filtering. As the data collection progresses, the PLS factors computed with the original calibration data are no longer optimal. A model updating strategy is applied here to revise the model.

On each prediction day, buffer spectra were collected before the sample collection. In the updating procedure, those buffer spectra were combined with the original calibration data to form a new calibration set for each prediction day. The buffer spectra are taken as “0” concentration calibration samples. Up to 48 buffer air-absorbance spectra were added. On some days, fewer buffer spectra were collected and were all added. The same parameter optimization was performed separately on different prediction days, because the calibration set has changed. Figure 5.13 depicts SEP values based on the optimization results corresponding to each prediction day, as well as the corresponding bias values. By using the updating strategy, not only



(A)

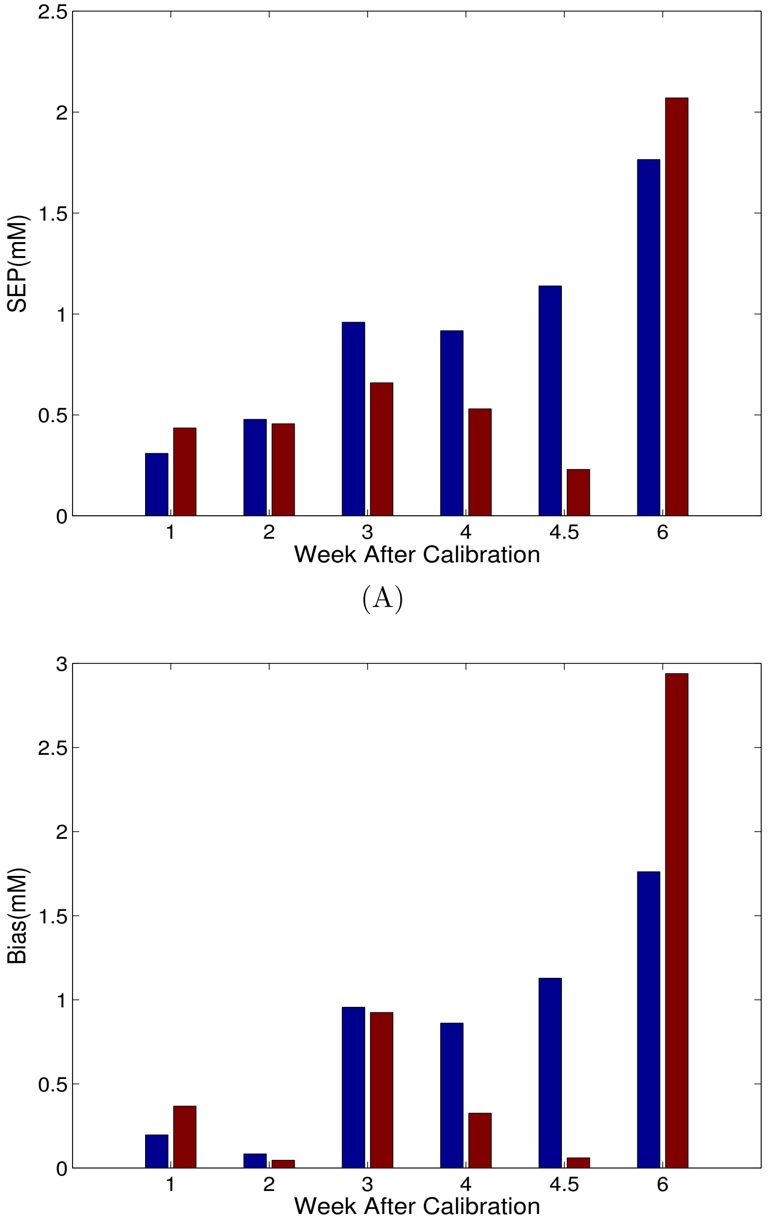
(B)

Figure 5.12. Comparison PCA score plots of all spectra before and after digital filtering. (A) Before digital filtering. (B) After digital filtering. Blue: Calibration. Red: Prediction 1. Green: Prediction 2. Purple: Prediction 3. Cyan: Prediction 4. Magenta: Prediction 5. Black: Prediction 6

the error of prediction has been improved, but also the consistent bias has been eliminated, especially on weeks 3, 4, and 4.5. However, for the last prediction data set, the updating was not helpful. This could be because of the buffer spectra collected on that day. From Figure 5.11A, the SEP value of this data set is over 4 mM when using buffer as the spectral background. The large value indicates that the buffer spectra did not perform well as backgrounds compared to the other data sets. When using buffer as background, it removes some of the solution information (e.g., water background). But in this data set, the expected effect is not evident. A similar PCA study to that described in Section 5.3.6 was done to compare the updating results. Figure 5.12 shows the score plot of the updated calibration set and the prediction set before and after the signal processing. It can be observed that after digital filtering, the first prediction set was blended well with the calibration data. But in the last prediction set, the calibration and prediction data are still separated. This underscores that the digital filtering step can help to reduce the effects of spectral variation with time but cannot overcome all such variation.

5.3.8 Comparison Results with PLS without Signal Processing

A grid search about the spectral range and the number of PLS factors was performed to find the best calibration model without the use of digital filtering preprocessing. In the grid search, the spectral range was selected over 4800 to 4200 cm^{-1}. The spectral range varied from 200 to 600 cm^{-1} with a 20 cm^{-1} increment. The moving step of the range was 5 cm^{-1}. Meanwhile, the number of factors was



(A)

(B)

Figure 5.13. Compared SEP and bias plot of the prediction results without and with the updating. (A) shows the SEP values. (B) is the absolute bias value. Blue bar: using air absorbance spectra; Red bar: with the updating strategy

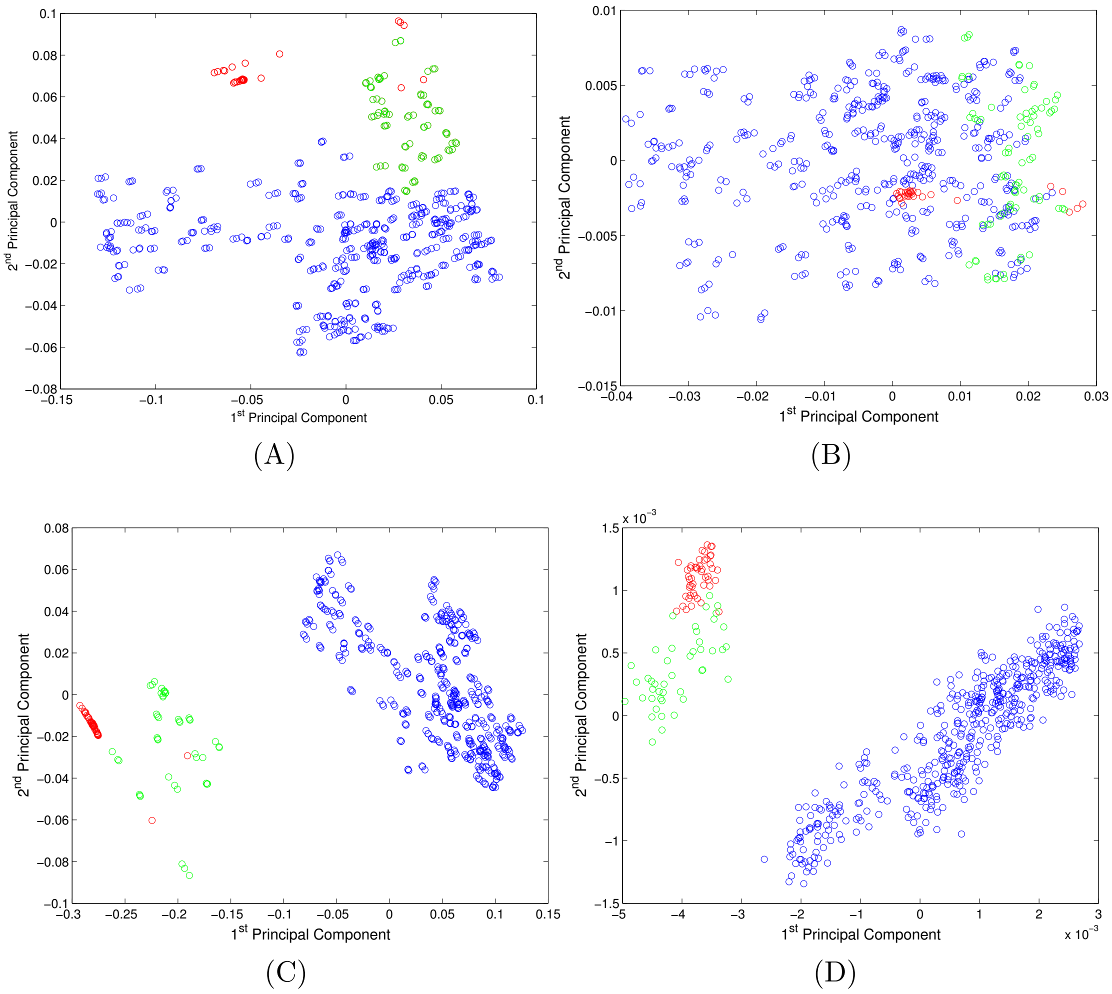
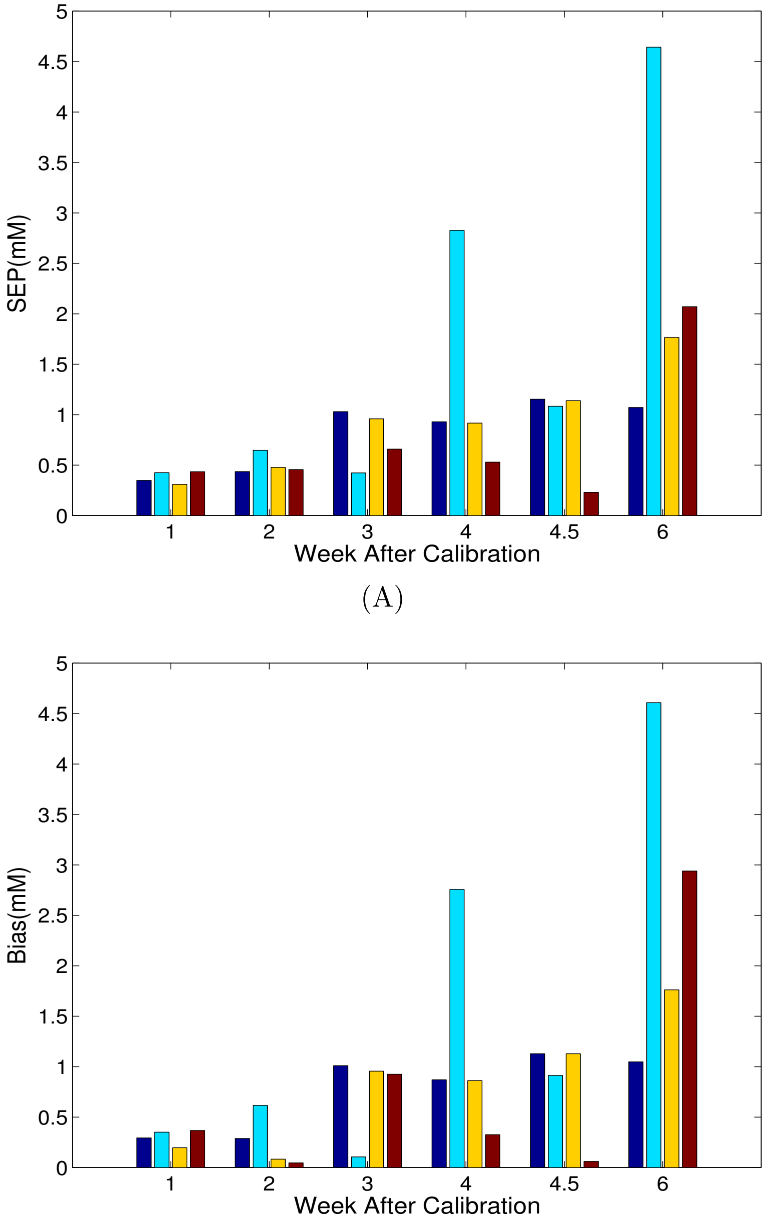


Figure 5.14. PCA score plots of the buffer updating results. (A) shows the result before the digital filtering of prediction set 1, which was collected a week after the calibration. (B) is the plot after digital filtering. (C) and (D) show the last prediction set before and after digital filtering, respectively. In the subfigures, blue circles represent the original calibration data, red denotes the buffer spectra used for updating the calibration model, and green indicates the prediction spectra.

also selected by the F-test. The maximum number of latent variables was 15. A leave-10%-out cross-validation with consecutive blocks was applied to find the optimal spectral range and latent variables. The updating method was also studied with PLS without digital filtering. Figure 5.15 shows the SEP and bias values of the compared methods. Inspection of Figure 5.15 reveals that the first two prediction sets perform similarly well with and without digital filtering and updating. Here, the calibration and prediction data are sufficiently close that no filtering or updating is beneficial. For the four later prediction sets, updating without filtering produces erratic results (e.g., the best results in week 3 but far and away the worst results in weeks 4 and 6. The static PLS model without filtering degrades by a factor of two in SEP for the last four weeks and does not compete well with the combination of filtering and updating in weeks 3, 4, and 4.5. Week 6 is the outlying case, as filtering and updating fail to perform well, while the static PLS model maintains its degraded performance (i.e., does not degrade further). Overall, however, the combination of filtering with updating proves to be the best among the data processing strategies tested.

5.4 Conclusions

The purpose of this study was to develop a filter design method to build calibration models for continuous monitoring in a four-component flowing system. The methodology was applied to both short- and long-term data collected with an AOTF-based spectrometer.



(A)

(B)

Figure 5.15. Compared SEP and bias plot of the prediction results without and with signal processing. (A) shows the SEP values. (B) is the absolute bias value. In both figures, the blue bars represent the PLS method, and cyan bars show the value produced by PLS combined with the updating method. Yellow and red bars represent the digital filtering results without and with updating, respectively.

Buffer spectra collected before the sample collection and air reference spectra collected simultaneously with the sample collection were compared for their utility as background spectra. From the SEP values and the corresponding correlation plots, for short-term prediction, there was no significant difference in background selection. However, it was found that the air reference spectra worked more stably and reliably than buffer spectra when applied to the long-term data (Group 2). While buffer spectra could only contain the solution information before sample collection, the air reference includes instrumental variation throughout the data collection.

A signal processing method was investigated to remove non-analyte variation which could affect the robustness of the calibration model. A Chebyshev type II bandpass filter was applied in a preprocessing step. With a PCA study performed on the long-term data, the score plots before and after the signal processing proved the positive effect of digital filtering. In quantitative analysis, an F-test after PLS regression was used to avoid over-fitting in calibration. The PSO method was employed in parameter optimization for filter design and PLS regression. This population-based optimization method helped reduce the large scale of the calculation and avoided the tedious process of a large-scale grid search.

Lastly, in order to stabilize the calibration model for long-term monitoring, a buffer updating strategy was studied. By adding the buffer spectra as calibration samples to the original calibration set, a new calibration set specified for each prediction day was defined and followed by the same data analysis procedure. This updating protocol helped improve prediction performance by lowering the bias between predicted and observed concentrations.

CHAPTER 6 SIGNAL PROCESSING METHODS FOR CONTINUOUS BIOREACTOR MONITORING WITH AUGMENTED CLASSICAL LEAST-SQUARES

6.1 Introduction

Process monitoring is an increasingly important topic in the field of monitoring and evaluation.[102–104] In the chemical and pharmaceutical industries, a good understanding of the manufacturing process is important in order to meet increasing demands of safety, yields and reduced waste. For example, in pharmaceutical applications, controlling and mastering the different process phases enables constant drug production quality, a decrease in the number of rejected batches and a shorter time to market.[105]

In the chemical industry, the use of batch or semi-batch processes has become popular to increase the adaptability of manufacturing facilities in constantly changing situations. High-quality and value-added specialty chemicals (e.g., polymers, biochemicals, food, semiconductors and agricultural chemicals) are commonly produced by batch processing methods.[105] In contrast to the operation of a continuous process, a facility used for batch processing must be continually stopped, started, and/or retooled each time the process is changed. This places extreme importance on the consistency and reproducibility of the process and makes effective process monitoring a critical requirement.

In both the chemical and pharmaceutical industries, analysis of key variables related to a manufacturing process often requires sophisticated chemical measurements rather than simpler and more traditional physical measurements such as temperature, flow rate, and pressure. The United States Food and Drug Administration has recognized the value of process chemical measurements through the establishment of their Process Analytical Technology (PAT) initiative.[106]

Process monitoring/analysis divides into five different classes according to the sampling and transportation of the analyzed sample. These analytical approaches are described as: (1) off-line, (2) at-line, (3) on-line, (4) in-line, and (5) noninvasive methods.[107]

Off-line analysis involves manual sampling with transport to a remote laboratory where techniques such as chromatography or mass spectrometry are applied. An at-line procedure is based on manual sampling with analysis conducted nearby (i.e., outside of a traditional laboratory). The analytical methodology employed is often relatively simple. Examples are chemical paper tests or titrations. On-line analysis is keyed by automated sampling and transportation to a nearby analyzer that can be run unattended. On-line gas chromatography or various spectroscopic methods can be employed. With an in-line procedure, the sampling interface is located in the process line itself and there is no need for a separated sampling system. For example, a pH probe or fiber optic spectroscopic probe could be inserted into a process pipe or reactor.[102] Finally, a noninvasive method can interrogate a process directly without having to come into physical contact with the sample. An example noninvasive method would be a reflectance measurement made through a window in a reaction vessel.

In this chapter, an in-line process analysis is implemented to monitor cell growth and protein production in a bioreactor. *Pichia pastoris*, a methylotrophic yeast, was employed. This type of cell has been successfully used in the expression of a variety of heterogonous proteins. Additionally, *Pichia pastoris* also carries the advantages of (1) ease of genetic manipulation and culture and (2) amenability to growth to high cell densities.[108–110].

The bioreactor run in this study is a three-stage process for the production of foreign proteins from *Pichia pastoris* cells. In the first stage, the engineered strain is batch cultured for biomass accumulation. This step uses a simple medium with glycerol as the commonly used carbon source. The second stage involves a glycerol feed transition phase. The biomass of the culture further increases at a limited growth rate, and the cells are prepared for protein production. After the desired cell density is achieved, the glycerol levels are allowed to deplete for a brief period and the cells are starved. At the third stage, the methanol feed begins. Methanol is added to the culture to induce the protein production, and the methanol concentration levels are maintained while the targeted protein is expressed.

To monitor the cell growth and the production of the protein, the concentration levels of the carbon sources are required to be controlled in a certain range. For example, the glycerol concentration must be above 0 to insure the reproduction of the cells, while the methanol concentration should be maintained between 3 g/L and 10 g/L during the third stage. A typical way to monitor the concentration is to take samples regularly (e.g., every hour) and obtain concentration estimates via a reference method. This is a tedious procedure, however, given that typical bioreactor runs can last for 3-8 days. Frequent monitoring requires a significant investment of labor resources, including night shift labor.

In the work described in this chapter, a spectroscopic method is developed to allow real-time continuous monitoring of the bioreactor process. As discussed previously, NIR spectroscopy is a nondestructive method. The lower background absorption from water in the NIR spectral region makes it possible to measure an aqueous solution without further sample preparation. A tubing circulation system can easily provide the sample introduction to the spectrometer and thereby enable a continuous spectral collection for real time monitoring. In terms of the spectrometer to use, a conventional laboratory Fourier transform (FT) IR spectrometer is expensive and difficult to minimize in size. Additionally, the moving mirror in the interferometer places limits on the ruggedness of the instrument for operation in an industrial environment. By contrast, an instrument such as the filter-based AOTF spectrometer described in Chapter 5 is potentially lower in cost and more compact than an FT design. The all solid-state components of the AOTF instrument provide for a rugged design. Meanwhile, the AOTF is capable of fast scanning over a relatively wide spectral range. The device and theory involved in the AOTF spectrometer have been introduced in Chapters 2 and 5.

A key element of effective process monitoring is the use of statistical methods for analysis of data and decision-making.[111] Multivariate analysis techniques from chemometrics can be applied to the analysis of the process data and can also assist in the improvement of the process. The techniques utilized typically consist of multivariate modeling methods such as principal component analysis (PCA) and partial least-squares (PLS) for use in building quantitative models for prediction of concentrations or properties related to the process being monitored.[112,113] In addition, techniques from statistical process control are used routinely for decision-making related to whether a process is within targeted specifications.[114]

When multivariate modeling techniques are applied to process monitoring, a key question is how to acquire the required data for use in building the desired models. The application of methods such as PCA or PLS to NIR spectroscopic data typically requires the performance of a set of initial calibration experiments to collect spectra and associated reference measurements for use in building models that will be subsequently applied to spectra collected in the future. The experiments and methodology described previously in Chapter 5 are representative of this process.

In the context of bioreactor monitoring, the analogous calibration measurements would require one or more bioreactor runs to be performed during which spectra are acquired and as many external reference measurements as possible are taken. Calibration models would be developed from the resulting data and then be available for application during subsequent bioreactor runs.

The critical element governing the success of this procedure is whether the developed calibration models are robust enough to work effectively with future data. The significant labor associated with the reference measurements dictates that the calibration procedure must not have to be performed very often. Two principal factors govern the degree of success that is attainable: (1) instrument stability and (2) process stability.

The resistance of the spectrometer to the effects of instrumental drift is extremely important in determining the overall robustness of the calibration model. As described in Chapter 5, even with the internal air reference, calibration models built from data collected with the AOTF spectrometer were observed to degrade over time.

Instability in the process itself is potentially even more problematic. A calibration model built with PCA or PLS is very heavily tied to the calibration data used in its development. As described previously in Chapter 3, PCA and PLS are empirical modeling methods that derive underlying factors or spectral shapes from the calibration data. The driving force in these methods is to explain the variance in the calibration spectra (PCA) or to jointly explain the variance in the calibration spectra and the covariance between the spectra and reference values (PLS). For the derived factors to be useful, both approaches have to be presented with calibration spectra that are representative of all future spectra to which the model will be applied. Any artifacts or unique elements that contribute significant variance to the calibration spectra will be encoded into the calibration model and may negatively impact its performance if these elements are not present in the future data to which the model will be applied.

Unfortunately, a bioreactor process is an inherently unstable system. During the course of the bioreactor run, the background sample matrix is constantly changing as the cells grow, consume nutrients, and produce waste. The need for pH adjustment or control of foaming may require additional reagents to be added. The starting media for the cell growth contains many components, some of which are derived from natural sources and may be subject to variation from batch to batch. Consequently, it is extremely difficult to guarantee that the calibration data collected from one run will be completely representative of future runs.

To address these limitations, this research explored the use of methodology that is less reliant on the collection of elaborate calibration data and less dependent on empirically derived spectral factors. Calibration models were constructed with augmented classical least squares (ACLS)[63] This method is an enhancement of traditional CLS spectral modeling[115], and makes use of pure-component spectra of known sample constituents, coupled with acquired background spectra to represent unknown components of the sample matrix. As introduced in Chapter 5, digital filtering will again be used to preprocess the acquired spectra to remove unwanted variance before submission to the calibration model. To address the dynamic nature of the bioreactor process, optimization of spectral and filter parameters will be used to tune the methodology over time to account for variation in the sample matrix from run to run. A novel synthetic data generation procedure is also introduced to obtain monitoring data for use in assessing model performance.

6.2 Experimental

6.2.1 Instruments and Apparatus

As in Chapter 5, a filter-based AOTF spectrometer was employed in the data acquisition phase of this study. Data collection was performed by our collaborators at ASL Analytical, Inc. Details about the instrumentation have been introduced previously in the discussion of Figure 5.2 and will be summarized briefly here. A 10-Watt tungsten lamp provided the light source. A beam splitter reflected 10% of the scanned light to the reference channel. Light from each channel was collected with two-stage thermo-electrically cooled InGaAs detectors. A single spectrum was based on 600 spectral sweeps within a 15-second integration time. The scanning wavelengths ranged between 5000 and 4000 cm^{-1}. Samples flowed continuously into the spectrometer from the bioreactor through a 1.3-mm inner diameter Teflon tube and spectra were scanned directly through the tubing. Teflon exhibits minimal absorption in the 4000 to 5000 cm^{-1} range.

6.2.2 Data Collection and Partitioning

The spectra employed in this work consisted of a combination of laboratory data designed to simulate the chemical composition of bioreactor samples and bioreactor data acquired during actual growth of *Pichia pastoris.* Each of these data sets will be described separately below.

6.2.2.1 Laboratory Data

The laboratory data were collected from mixture samples containing varying amounts of glycerol, methanol, and sorbitol in a phosphate-buffered aqueous matrix that simulated the typical growth media used in *Pichia* bioreactors. Mixture samples were generated in real time by use of an in-house constructed system of computer-controlled peristaltic pumps. Stock solutions of glycerol (40.0 g/L, Sigma Chemical Co., St. Louis, MO, >99.0%), methanol (10.0 g/L, Fisher Scientific, Pittsburgh, PA, ACS reagent), and D-sorbitol (18.2 g/L, Sigma, ≥98%) were prepared in the background matrix and combined via the pumping system to produce mixture samples in which the glycerol, methanol, and sorbitol concentrations ranged from zero to the stock solution concentrations noted above. The background media contained yeast extract (∼10 g/L, Sensient Bio-Ingredients, Milwaukee, WI), yeast nitrogen base (YNB) (∼13 g/L, Becton Dickinson Co., Franklin Lakes, NJ), soy peptone (∼20 g/L, Kerry Bio-Science, Norwich, NY), potassium phosphate dibasic (∼2 g/L, Sigma, ACS reagent), and potassium phosphate monobasic (∼12 g/L, Fisher, ACS reagent). In addition, 5-fluorouracil (∼0.4 g/L, Sigma, 99%) was added as a preservative.

The open circles in Figure 6.1A denote the 130 pure and mixture samples generated through this approach. The sample collection protocol consisted of 14 repeats of the following pattern: (1) background matrix, (2) glycerol stock solution, (3) methanol stock solution, (4) sorbitol stock solution, (5) five mixture samples. The background matrix and the three stock solutions were then measured one final time to complete the run. After setting the appropriate pump speeds to produce a given sample, the solution was flowed through the spectrometer for 7 min. The complete run of 130 samples required approximately 15 h and was performed in an automated manner with only periodic inspection. This experiment was performed three times over 17 days (days 1, 16, 17).

Each peristaltic pump was calibrated at the beginning and end of each block of 130 samples. Pump calibration consisted of delivery of water into a receiving vessel placed on an analytical balance. The resulting mass vs. time data were used to correlate rotor speed with observed pump flow rate. For the mixture samples, component concentrations were then computed by use of the calibrated flow rates and stock solution concentrations. Concentrations of individual samples were taken as the average of the concentrations produced by the two pump calibrations. Pooled standard deviations in the glycerol, methanol, and sorbitol concentrations computed across the pump calibrations were less than 0.03 g/L except for glycerol on day 1 (0.15 g/L).

As described in Chapter 5, one complication that affected the data was the occurrence of air bubbles in the sampling tube as it flowed through the spectrometer. Air bubbles are a problem in the 4000-5000 cm^{-1} spectral region because they result in a reduction of absorbance of the aqueous background and thus cause a change in the shape of the single-beam spectrum. The data acquisition software controlling the spectrometer was designed to detect the presence of air bubbles through this change in the spectral shape. Any spectra that were determined to have air bubbles were discarded and not stored. For the run on day 1, air bubbles were a significant problem and resulted in a loss of data for a number of samples. In addition, the gaps in the data resulted in difficulty in reliably assigning spectra to concentrations. For this reason, spectra of only 82 of the 130 samples on day 1 could be assigned to reference concentrations. These samples are indicated by the hatched circles in Figure 6.1A. For days 16 and 17, all 130 samples were recorded.

6.2.2.2 Bioreactor Data

As introduced previously, each bioreactor run involves three steps, the glycerol feed for cell culturing, the transition phase in which the cells are starved, and the protein production step with methanol as the carbon source. Specific experimental conditions may change from run to run. The research presented here was based on five individual bioreactor runs (Runs 1-5) performed at the Center for Biocatalysis and Bioprocessing at the University of Iowa. For simplicity in describing the experimental protocol, the details of Run 1 are provided below as an example of typical conditions.

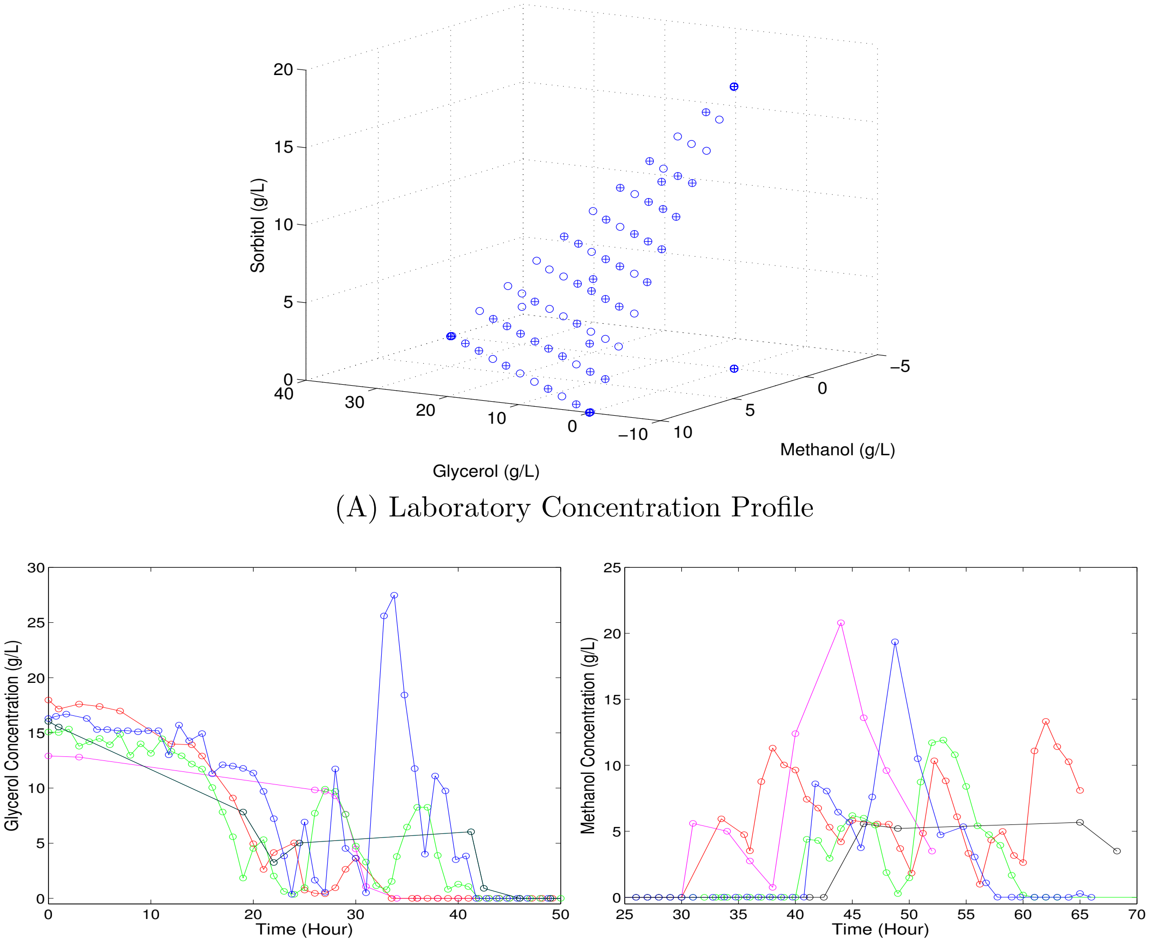
At the time the reactor was inoculated with *Pichia* cells, the total liquid volume of the fermentor was ∼8 L with around 110 g glycerol. Other major species for the initial cell growth state included 35 g ammonium sulfate, 1.75 g YNB, 78 g spray-dried corn steep, 0.4 mg biotin, and 5 mL anti-foaming reagent. Specific sources for these reagents were not available to us. The pH of the liquid was controlled to 5.0 with ammonium hydroxide. The initial reactor temperature was 30 °C. The glycerol feed stage lasted ∼40 h. When the glycerol concentration dropped to ∼0 g/L, the methanol feed for the protein production was started by adding 40 g of methanol (~50 mL), 7 g YNB, 2 mg biotin, and 50 mL deionized water. The protein production lasted ~30 h.

After the inoculation, the concentrations of the analytes of interest (glycerol and methanol) were monitored by both the spectroscopic and reference methods. The liquid in the fermentor was pumped out and flowed through a circulated Teflon tube and introduced to the AOTF spectrometer. Spectra were recorded continuously at 15 s intervals. To obtain the reference concentrations, samples were collected hourly and analyzed for glycerol and methanol concentration by use of an Analox GM8 Multi-Assay Analyzer (Analox Instruments, Ltd., London, UK). This analyzer uses a Clark-type amperometric oxygen electrode to monitor the enzyme-catalyzed oxidation of the selected analyte. The concentration profiles obtained from the reference method for glycerol and methanol are plotted as a function of time in Figures 6.1B and 6.1C. In the figure, circles represent the concentration points collected by the reference method. The points are connected by lines without interpolation.

6.2.2.3 Data Partitioning

In the analysis of the laboratory data, the 82 samples collected on Day 1 were used as calibration data for both glycerol and methanol. Model validation was performed on the rest of the samples from Days 16 and 17. For each sample, the three spectra located at the center of the block of spectra defining the concentration were taken as replicates. No co-addition was performed.

For the bioreactor data, the starting conditions of each run were different in



(A) Laboratory Concentration Profile

(B) Bioreactor Profile of Glycerol

(C) Bioreactor Profile of Methanol

Figure 6.1. Concentration profiles. (A) Experimental design of laboratory data collection for 130 mixture samples of glycerol, methanol, and sorbitol (open circles). Duplicate samples exist for the background matrix and pure-component solutions. Because of loss of data arising from the presence of air bubbles, only 82 samples on day 1 could be accurately assigned to spectra. These samples are indicated by the hatched circles. (B) Glycerol concentration profile of the bioreactor data from Runs 1 to 5 obtained from the reference measurements. In each figure, circles represent the collected reference samples. Samples within each run are connected by straight lines without interpolation. (C) Methanol concentration profile for bioreactor runs. The format of the presentation is the same as in the panel above. Run 1: Magenta. Run 2: Red. Run 3: Green. Run 4: Blue. Run 5: Black.

terms of the feed protocols (i.e., the background sample matrix) and temperatures. For this reason, calibration models were built individually for each run. During the cell culturing stage, spectra from the first 5 hours after the inoculation were used in building the glycerol model. Spectra collected 5 hours before the start of the methanol feed were employed in methanol modeling. Details about the modeling procedures will be presented in the next section.

6.3 Data Analysis Strategy

6.3.1 Overview of Data Analysis

As discussed previously, the AOTF spectrometer used in this research provides separate sample and reference detector channels. Accordingly, each collected sample spectrum has a corresponding air reference as the background and air-absorbance spectra were used in all further calculations. Spectra were deresolved from 0.67 cm^{-1} point spacing to 4 cm^{-1} to reduce the load of the calculations. Spectra were preprocessed as described previously in Chapter 5. Chebyshev Type II bandpass digital filtering was applied followed by submission of the filtered data to the calibration model calculation. Parameter optimization was again performed with particle swarm optimization (PSO).

Because of the limited reference data and the dynamic nature of the bioreactor sample matrix, calibration models were computed with the ACLS method rather than with the PLS modeling technique used in Chapter 5. It was hypothesized that insufficient reference data were available to assemble a set of globally representative calibration data for use in computing a reliable PLS model that could perform well in future predictions. The ACLS method is less reliant on the collection of large quantities of calibration data and is amenable to an implementation in which a new calibration model could be computed for use with each bioreactor run.

As implemented in Chapter 5, the PSO optimization of the filtering and modeling parameters requires a set of monitoring data for use in assessing the performance of each parameter set evaluated. Because of the limited amount of reference data, a synthetic monitoring set was established based on the linearly additive property of the Beer-Lambert law. The generated synthetic monitoring data were also evaluated for use in constructing PLS models. The concept of using synthetic NIR spectra to build calibration models has been evaluated in previous work[84,116], as well as in the infrared remote sensing research described in Chapter 4.

6.3.2 Synthesis of Monitoring Spectra for Bioreactor Data

Before the inoculation of the reactor with *Pichia* cells, spectra are typically acquired of the base media initially added to the fermentor. This media contain a fixed and known amount of glycerol and is also known to contain no methanol. Thus, spectra acquired before inoculation can be used as representative of the experimental conditions and starting sample matrix associated with the current bioreactor run. Examination of the reference data obtained for the six bioreactor runs also suggests that during the first five hours after inoculation, the glycerol concentration remained relatively constant because the initial cell growth and corresponding consumption of glycerol was very slow. Data collected during this period can also be used to extract background information.

Due to the linear relationship between concentration and absorbance encoded in the Beer-Lambert law, the glycerol contribution to the measured absorbance can be subtracted based on the known glycerol concentration and its corresponding pure-component absorbance spectrum. To a close approximation, the resulting spectra can be considered glycerol-free and representative of the background sample matrix. These spectra were subsequently used to generate the monitoring data.

The spectra in the background matrix were obtained by mathematical calculation as in Eq. 6.1, where $A_{\text{Glycerol-free},i}$ is the absorbance in the obtained background spectrum at point i, A_i is the measured absorbance, c_{Initial} is the known initial glycerol concentration and $A_{\text{Unit-Glycerol},i}$ is the collected pure-component spectrum at unit concentration.

$$A_{\text{Glycerol-free},i} = A_i - c_{\text{Initial}} \times A_{\text{Unit-Glycerol},i} \tag{6.1}$$

Before each run, pure-component spectra of glycerol and methanol with known concentrations were collected, as well as a spectrum of water. The pure-component spectra at unit concentration were estimated by Eq. 6.2. In the equation, $A_{\text{Pure-Analyte},i}$ is the absorbance of the pure-component spectrum of glycerol or methanol (concentration of c_{Analyte}) at point i and $A_{\text{Water},i}$ is the absorbance of pure water. The calculation in Eq. 6.2 is only approximate as it assumes that the spectra scale linearly with concentration and that changes in glycerol or methanol concentration do not appreciably change the water background spectrum by solvent displacement effects.

$$A_{\text{Unit}-\text{Analyte},i} = \frac{A_{\text{Pure}-\text{Analyte},i} - A_{\text{Water},i}}{c_{\text{Analyte}}} \tag{6.2}$$

After the calculations in Eqs. 6.1-6.2, the spectra were ready to be used in generating the spectra in the monitoring set for the glycerol model. For the monitoring set used in optimizing the methanol model, before methanol was added, all spectra were methanol-free and could be used as backgrounds. Spectra collected five hours before adding methanol were used to provide the best match in backgrounds to the methanol-containing spectra collected subsequently. Spectra collected during this time period were also always low in glycerol concentration and did not require any mathematical subtraction of the glycerol signature.

Every eight background spectra across the five-hour window were averaged to obtain around 250 spectra for use in generating synthetic data. To generate the monitoring data set, 200 of the background spectra were randomly selected and random amounts of the analyte were added in the range of 0 to 15 g/L. This procedure assumed linear additivity of the spectra and is described by Eq. 6.3. In the equation, $A_{\text{Monitoring},i}$ is the obtained absorbance of the monitoring spectrum at point i, $A_{\text{Background},i}$ is the absorbance for the background spectrum, and c_{Analyte} and $A_{\text{Unit}-\text{Analytie},i}$ are as described previously.

$$A_{\text{Monitoring},i} = A_{\text{Background},i} + c_{\text{Analyte}} \times A_{\text{Unit}-\text{Analyte},i} \tag{6.3}$$

6.3.3 Study of Regression Methods

As mentioned above, the limited number of reference points and varied initial experimental conditions made it difficult to use the PLS method to develop calibration models. This led to the use of the ACLS method in this work.

For chemical systems at low concentrations, the absorbance spectrum of a sample can be approximated as the sum of absorbances from each constituent species. As introduced in Chapter 3, with the knowledge of each pure spectrum and the composition of the chemical system, a CLS model can be built. For a matrix, $\mathbf{A}$, containing n absorbance spectra (p spectral points) in the columns, a typical CLS model is written as:

$$\mathbf{A} = \mathbf{KC} + \mathbf{E} \tag{6.4}$$

where $\mathbf{K}$ is a matrix whose columns contain pure-component absorbance spectra at unit concentration and constant path length for each of h components, $\mathbf{C}$ is a matrix whose n columns specify the h component concentrations contributing to the spectra in $\mathbf{A}$, and $\mathbf{E}$ is the residual error matrix. For a simple chemical system, the $\mathbf{K}$ matrix can be assembled by obtaining the pure-component spectra of each species at unit concentration, and the concentration matrix $\hat{\mathbf{C}}$ can estimated

$$\hat{\mathbf{C}} = (\mathbf{K}'\mathbf{K})^{-\mathbf{1}}\mathbf{K}'\mathbf{A} \tag{6.5}$$

However, for the complicated bioreactor runs, especially considering the presence of the cells and the dynamic nature of the spectral background, it is impossible to specify **K** in Eq. 6.5 precisely. In this case, ACLS can be a useful method. The ACLS method is based on the recognition that the **K** matrix used in the concentration estimation in Eq. 6.5 can contain other "spectral shapes" in addition to the known pure-component spectra. All components except the target analyte can be considered as background species. For example, in the first half of the bioreactor runs, glycerol is the analyte of interest. All of the other species, including the cells, YNB, peptone, and water are part of an integral background. Their spectral information combines to define the background absorbance. Because the glycerol contribution in absorbance has, to a first approximation, been removed mathematically through the use of Eqs. 6.1-6.3, the five-hour block of background data can be used to extract the components or spectral shapes that define the underlying absorbance.

Principal component analysis (PCA)[3,61] was employed to extract a mathematical representation of the underlying components of the background. While the individual spectral loadings that define the principal components (PCs) do not correspond to the pure-component spectra of specific chemical species, together they provide an efficient representation of the background spectral variance. The incorporation of PCs into CLS models was introduced by Martens and Naes[61] and has been used by Haaland[63] in developing various ACLS methods. Olesberg et al. have used the approach to characterize the spectral background in *in vivo* NIR measurements of rat tissue.[94]

With the m computed spectral loadings as the estimation of the background matrix, the original CLS model in Eq. 6.4 can be modified as

$$\mathbf{A} = \mathbf{KC} + \mathbf{TP} + \mathbf{E} = \mathbf{K_a C_a} + \mathbf{E} \tag{6.6}$$

The terms **A, KC**, and **E** are as defined in Eq. 6.4, **P** is the (p×m) matrix of spectral loadings, and **T** is the (m×n) matrix of scores that encodes the contributions of the PCs in **P** in establishing the background for each of the *n* spectra comprising **A**. The matrices, **P** and **T**, are obtained from application of PCA to the five-hour block of background spectra. The model can be further defined by adding the background spectral loadings as additional columns in **K** to form the augmented $\mathbf{K_a}$ matrix. The concentrations in unknown spectra can then be estimated from Eq. 6.5 by using $\mathbf{K_a}$ instead of **K**.

In building the model for glycerol, the unit pure-component spectra calculated from Eq. 6.2 were used. No methanol was added in the model because it was not involved in the experiment during this period. The first two PCs resolved from the glycerol-free background spectra were then added into the $\mathbf{K_a}$ matrix with a linear baseline correction term. The methanol model was defined similarly. Because glycerol was still present in the background matrix used with the methanol model, besides the methanol pure-component spectrum, the glycerol pure-component spectrum was also added into $\mathbf{K_a}$. The spectral range and digital filtering parameters were studied in the optimization step with the simulated monitoring spectra.

For the laboratory data without cells present, spectra with only the background matrix flowing were collected to represent the background. Besides glycerol, methanol and sorbitol, pure-component spectra of YNB, peptone and yeast extract were also collected and could be added into the ACLS model. Here, the optimization also involved the selection of which additional components to add to the $\mathbf{K_a}$ matrix. This procedure will be introduced in the optimization section.

6.3.4 Optimization

Similar to the work described in Chapter 5, the optimization of the calibration model employed particle swarm optimization (PSO). We have previously discussed the details regarding the implementation of PSO. Similar procedures were used here.

In the execution of the optimization for the bioreactor data, the spectral range selection in the ACLS model and the filter parameters of the Chebyshev type II bandpass filter were studied. The spectral range was selected between 4800 cm^{-1} and 4200 cm^{-1} in 10 cm^{-1} intervals. The filter order was selected between 1 and 6, with the stopband attenuation ranging from 20 to 100 dB with an increment at 5 dB. The normalized frequency cut-off was studied between 0.01 and 0.99 with 0.01 spacing. Six parameters were optimized; thus, the dimension of the particles was six. The fitness function which guided the model convergence to the global best was the standard error in concentration predictions for the the synthetic monitoring set. This will be termed the standard error of monitoring (SEM). As in Chapter 5, the optimization was performed 10 times, and the model with the lowest SEM was selected for use in prediction. The optimized model was used without change throughout a given bioreactor run, but was repeated for each separate run.

In the laboratory non-cell experiment, pure-component spectra of glycerol, methanol, sorbitol, YNB, and peptone were also collected. Besides the analyte of interest, the pure-component spectra of other species, such as sorbitol and peptone, can also be added into the **K** matrix, or included in the background spectral loadings. Therefore, the optimization also involved a selection of the pure components used in the ACLS model. The extra spectra were added into the optimization parameters and controlled in a switch mode. In operation of the PSO, by randomly selecting either 1 or 0, the component was on or off in the model. With the component selection, there were 9 parameters in the optimization, including two parameters of the spectral segment, the number of latent variables, the filter order, the frequency cutoff (starting and ending points), the stopband attenuation, and three binary numbers which represented the additional components. The model which provided the lowest error value in predicting the calibration samples from the first run was chosen as the best model after 10 runs of the optimization with different initial populations of 50 particles.

Because the laboratory data contained a conventional set of calibration data with known concentrations, PLS models could also be computed for comparison with the models based on ACLS. Optimization of the spectral range and filter parameters was performed as outlined above. The maximum number of latent variables was 10 followed by the F-test as described in Chapter 5. Instead of three random draws of the monitoring set as done in Chapter 5, a leave-10%-out cross-validation procedure was employed to obtain a monitoring error as the fitness value. Each prediction subset in cross-validation was selected by consecutive blocks. The global optimum was selected based on the lowest cross-validated standard error of prediction (CVSEP) value obtained across 10 PSO runs. Each run had a different starting population with 30 initial sets of parameters. This model was selected for application to the two prediction sets.

6.4 Results and Discussion

6.4.1 Results for Laboratory Data

The optimization results in terms of the filter design and regression model for the non-cell laboratory data are shown in Table 6.1a for glycerol and 6.1b for methanol. Three data sets are considered: calibration data and prediction sets 1 and 2. The corresponding standard error of calibration (SEC) and standard error of prediction (SEP) values are listed in Tables 6.2a and 6.2b. In the figures and tables, ACLS_1 denotes the CLS model based on pure-component spectra of glycerol, methanol, sorbitol, and one spectrum representing the background matrix. The ACLS_2 model included the possibility of adding pure-component spectra of soy peptone, yeast extract, or YNB in addition to the components of the ACLS_1 model. Based on the prediction ability of the calibration and prediction sets, although the PLS and ACLS_2 models were generally better performing with the calibration data, the ACLS_1 method provided lower SEP values (below 0.5 g/L) while offering a competent SEC. Correlation plots are shown in Figures 6.2 and 6.3. Obvious bias in the prediction can be observed from both the $ACLS_2$ and PLS models. The PLS model is particularly bad in this respect.

The PLS method performs the prediction based on spectral factors (latent variables) extracted from the calibration data. Bias in the prediction results suggest the occurrence of some new source of variation in the intervening time between the calibration and prediction data. This led to the use of an updating strategy in Chapter 5. By contrast, after getting the optimized filter parameters, the CLS method uses the latest pure-component spectra and the background spectrum of the current run in assembling the $\mathbf{K}$ matrix. By automatically including data from the current prediction day, this approach potentially avoids the variation brought about by instrumental drift. However, if new components are introduced into the system, a repeat of the optimization of the filter and spectral range parameters is still needed.

6.4.2 Process Monitoring Results with Bioreactor Data

The optimization results describing the digital filters and spectral range parameters used with the ACLS models are listed in Table 6.3. Figure 6.4 depicts the process monitoring results for glycerol in the five bioreactor runs. The monitoring period covered from 5 hours after the inoculation started to the starvation period before the methanol was added. The blue crosses represent the predicted concentrations estimated by the spectroscopic method, while the solid red dots are the reference values. By matching the spectral acquisition time to the time when the reference sample was collected, the corresponding sample spectrum can be obtained. Therefore, the

Table 6.1. Filter design and calibration models for laboratory non-cell data

(a) Glycerol

Model	Filter Parameter			Regression Model	
	Order	Frequency Cut-off (Normalized)	Stopband Attenuation (dB)	Spectral Range (cm^{-1})	Note
ACLS_1	4	0.02 - 0.99	25	4700 - 4320	
ACLS_2	5	0.01 - 0.19	35	4695 - 4325	Peptone[a]
PLS	1	0.01 - 0.22	25	4760 - 4180	LV[b] = 9

(b) Methanol

Model	Filter Parameter			Regression Model	
	Order	Frequency Cut-off (Normalized)	Stopband Attenuation (dB)	Spectral Range (cm^{-1})	Note
ACLS_1	6	0.03 - 0.59	85	4805 - 4460	
ACLS_2	6	0.02 - 0.97	50	4745 - 4485	none[c]
PLS	2	0.01 - 0.99	35	4800 - 4300	LV[b] = 3

[a] Model included the pure-component spectrum of soy peptone.

[b] Number of latent variables in PLS model.

[c] Optimized model did not include additional pure-component spectra.

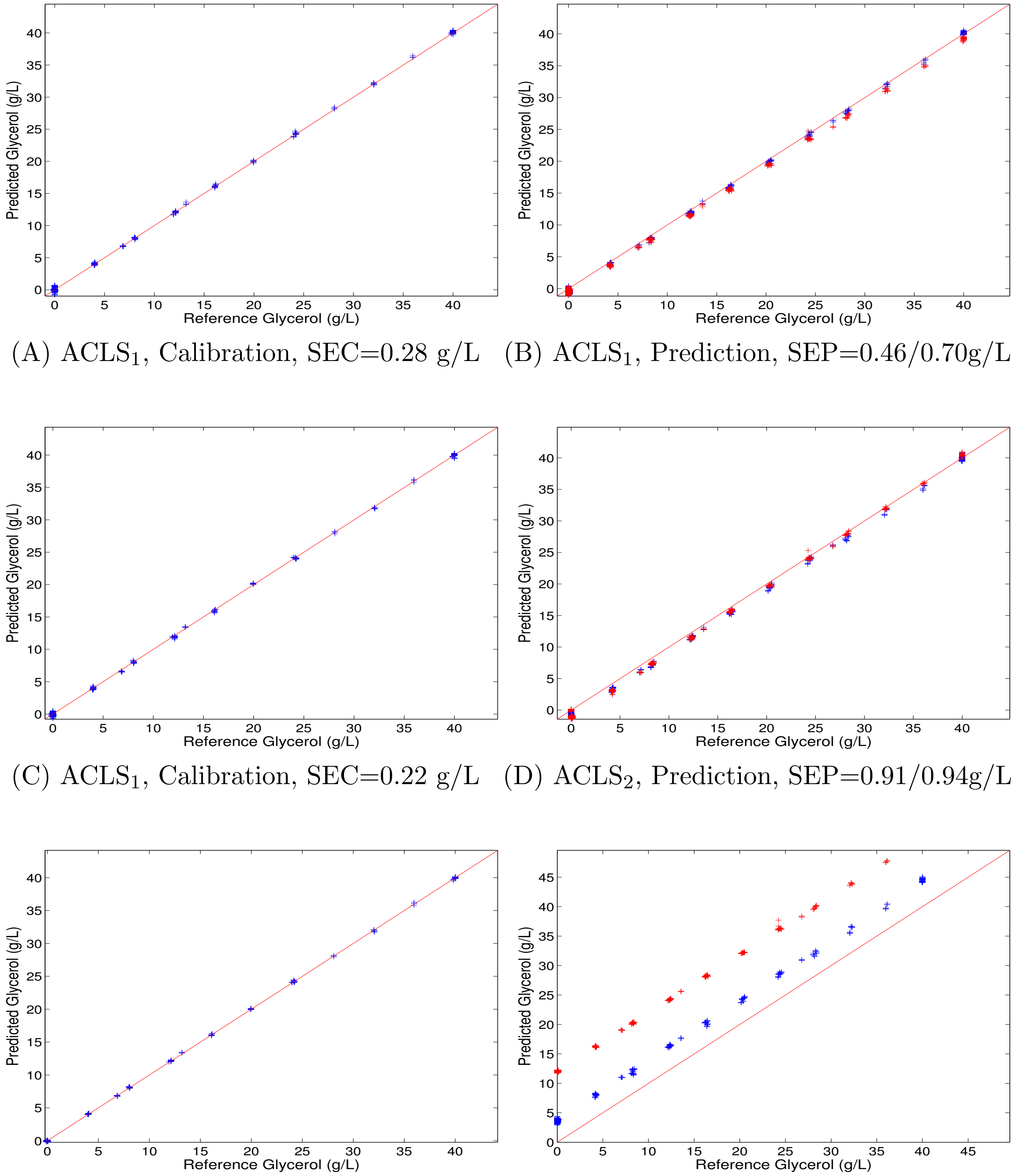
Table 6.2. SEP values (g/L) for non-cell data monitoring

(a) Glycerol

Data Set	ACLS_1	ACLS_2	PLS
Calibration	0.28	0.22	0.10
Prediction 1	0.46	0.91	3.9
Prediction 2	0.70	0.94	11.9

(b) Methanol

Data Set	ACLS_1	ACLS_2	PLS
Calibration	0.51	0.50	0.031
Prediction 1	0.70	0.92	0.77
Prediction 2	0.66	1.66	0.28

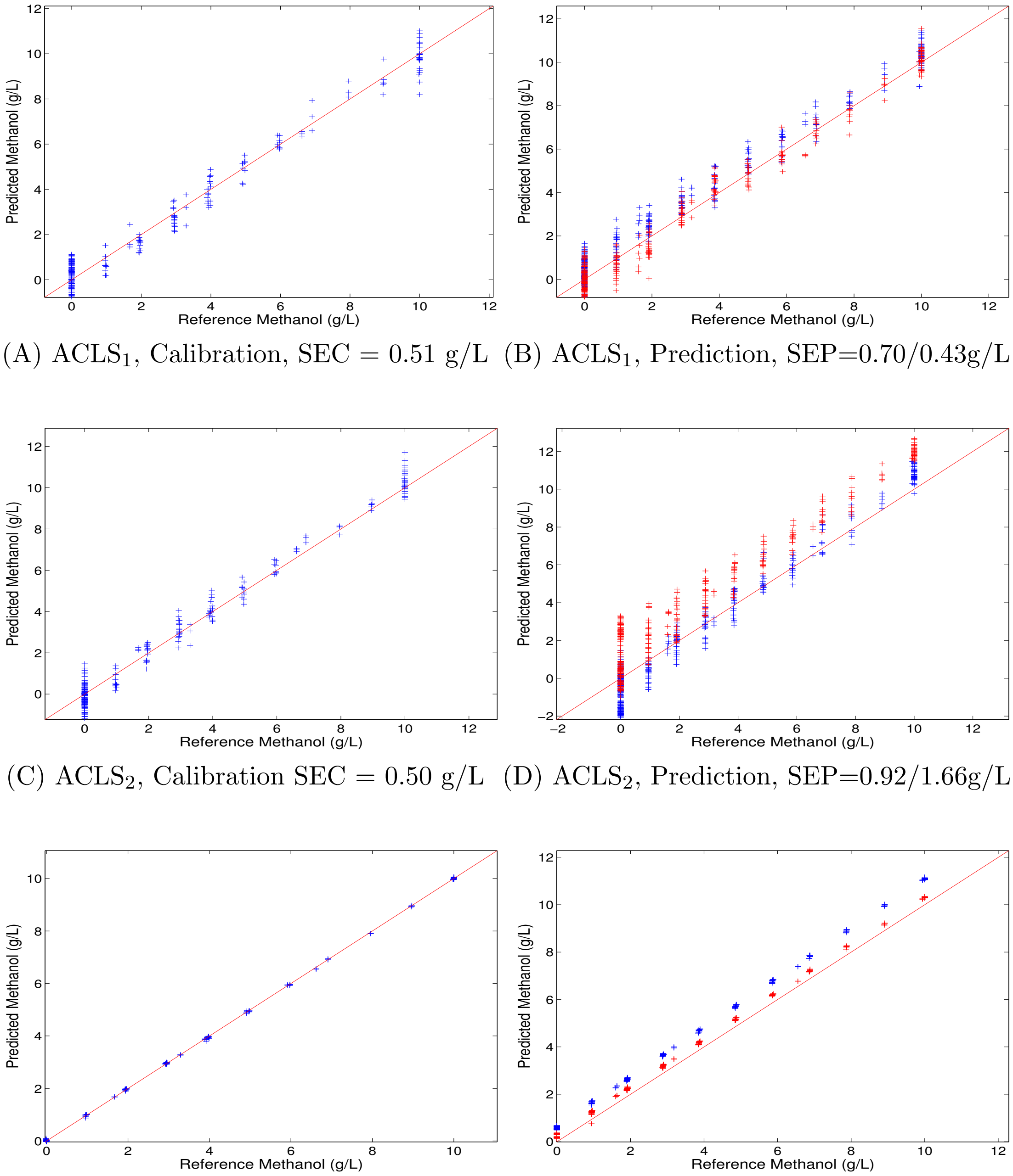


(A) ACLS$_1$, Calibration, SEC=0.28 g/L (B) ACLS$_1$, Prediction, SEP=0.46/0.70g/L

(C) ACLS$_1$, Calibration, SEC=0.22 g/L (D) ACLS$_2$, Prediction, SEP=0.91/0.94g/L

(E) PLS, Calibration, SEC = 0.10 g/L (F) PLS, Prediction, SEP = 3.9/11.9 g/L

Figure 6.2. Correlation plots for glycerol predictions in the laboratory data. The subfigure caption shows the corresponding algorithm, data set and error value (SEC or SEP). The ACLS models are termed ACLS$_1$ and ACLS$_2$. The ACLS$_1$ model used the three major components (glycerol, methanol and sorbitol) and the background in assembling the $\mathbf{K_a}$ matrix. In the ACLS$_2$ model, besides the major components and background spectra, the optimization involved a component selection from soy peptone, YNB and yeast extract. In the prediction plots, the blue and red symbols denote prediction sets 1 and 2, respectively.



(A) ACLS$_1$, Calibration, SEC = 0.51 g/L (B) ACLS$_1$, Prediction, SEP=0.70/0.43g/L

(C) ACLS$_2$, Calibration SEC = 0.50 g/L (D) ACLS$_2$, Prediction, SEP=0.92/1.66g/L

(E) PLS, Calibration, SEC = 0.031 g/L (F) PLS, Prediction, SEP = 0.77/0.28 g/L

Figure 6.3. Correlation plots for methanol prediction of the laboratory data. The subfigure caption shows the corresponding algorithm, data set and error value. The plot format is the same as in Figure 6.2

SEP value of each run can also be calculated and is listed in Table 6.5. In calculating the SEP values for the glycerol prediction, the time period was selected from five hours after inoculation until methanol was added.

It can be observed that, for some of the runs, the initial glycerol concentration has a larger variation to the calculated value. This was because the reference method obtained with the Analox instrument does not have a good response to high concentration samples. However, once the cells start to consume glycerol, the spectroscopic method provides a good consistency with the reference method. For Run 5, the large fluctuation at around 40 h is caused by a blockage of cells in the Teflon sampling tube used to move material to the spectrometer.

For methanol predictions, the optimization results for the digital filters and spectral ranges used in the ACLS models are listed in Table 6.4 and the prediction results are plotted in Figure 6.5. The time period monitored begins with the methanol introduction and continues until the end of the reference sample collection.

The SEP values for both methanol and glycerol predictions are listed in Table 6.5. Overall, methanol predictions are good except for Run 5 in which a large bias in predicted values exists.

6.4.3 Comparison Result with Methods without Signal Processing

Results obtained from this work are compared with the algorithms without signal processing. For the laboratory data, the spectral range was selected by a grid search on 4800 to 4200 cm^{-1} with the moving step at 5 cm^{-1}. The range size was

Table 6.3. Optimization results for glycerol predictions in bioreactor runs

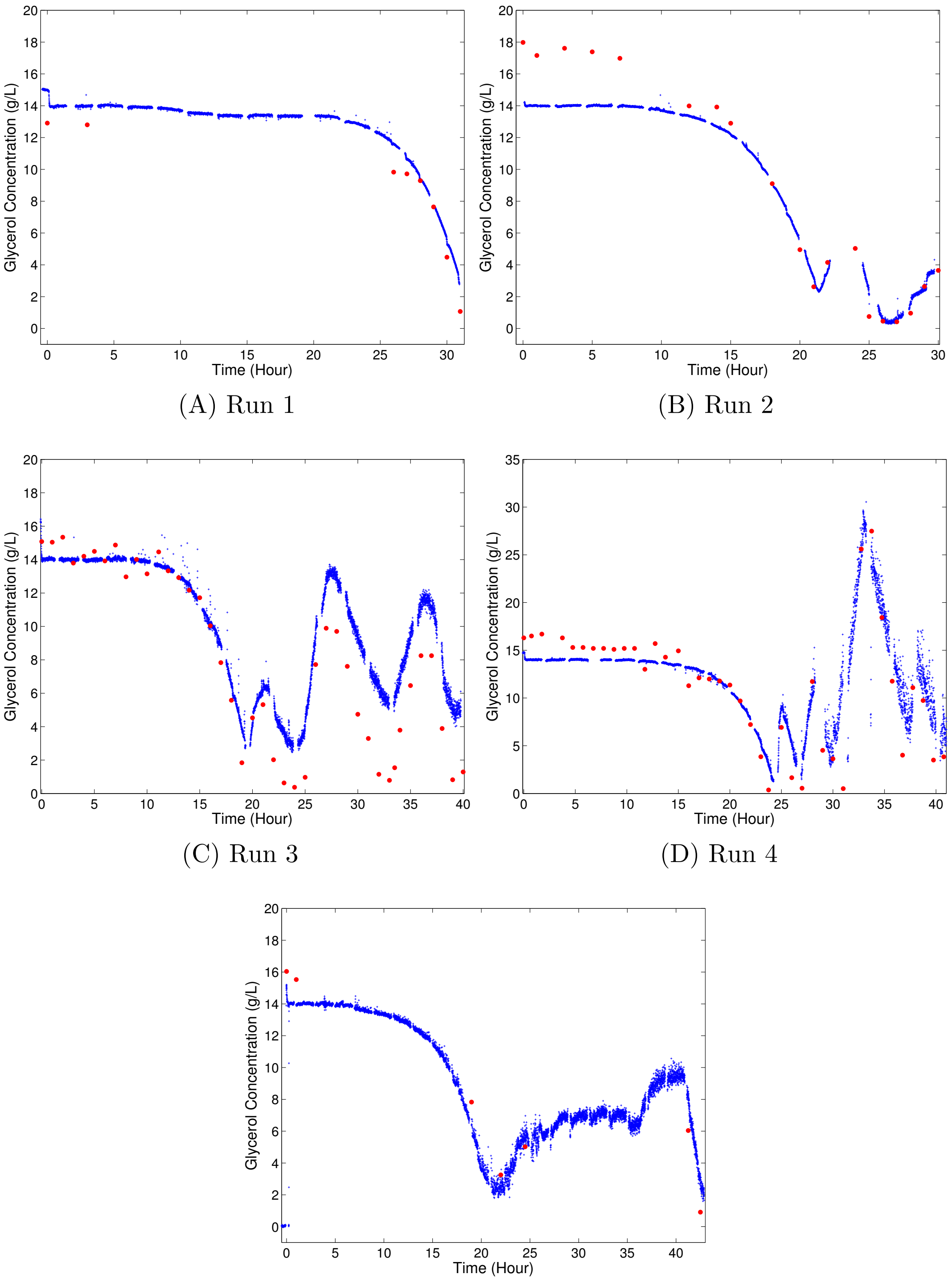
Model	Filter Parameter			ACLS
	Order	Frequency Cut-off (Normalized)	Stopband Attenuation (dB)	Spectral Range (cm^{-1})
1	6	0.01 – 0.49	20	4670 – 4220
2	2	0.01 – 0.21	25	4675 – 4270
3	1	0.01 – 0.63	95	4555 – 4350
4	2	0.01 – 0.50	30	4620 – 4380
5	2	0.01 – 0.56	35	4525 – 4310

Table 6.4. Optimization results for methanol predictions in bioreactor runs

Model	Filter Parameter			ACLS
	Order	Frequency Cut-off (Normalized)	Stopband Attenuation (dB)	Spectral Range (cm^{-1})
1	3	0.01 – 0.15	45	4620 – 4270
2	3	0.02 – 0.13	45	4760 – 4240
3	6	0.01 – 0.99	20	4700 – 4235
4	2	0.02 – 0.17	20	4620 – 4380
5	4	0.01 – 0.63	20	4730 – 4385

Table 6.5. Prediction results for bioreactor monitoring

Run	Glycerol SEP (g/L)	Methanol SEP (g/L)
1	0.98	2.84
2	1.24	1.28
3	2.71	0.91
4	2.14	3.42
5	1.32	0.23



(A) Run 1

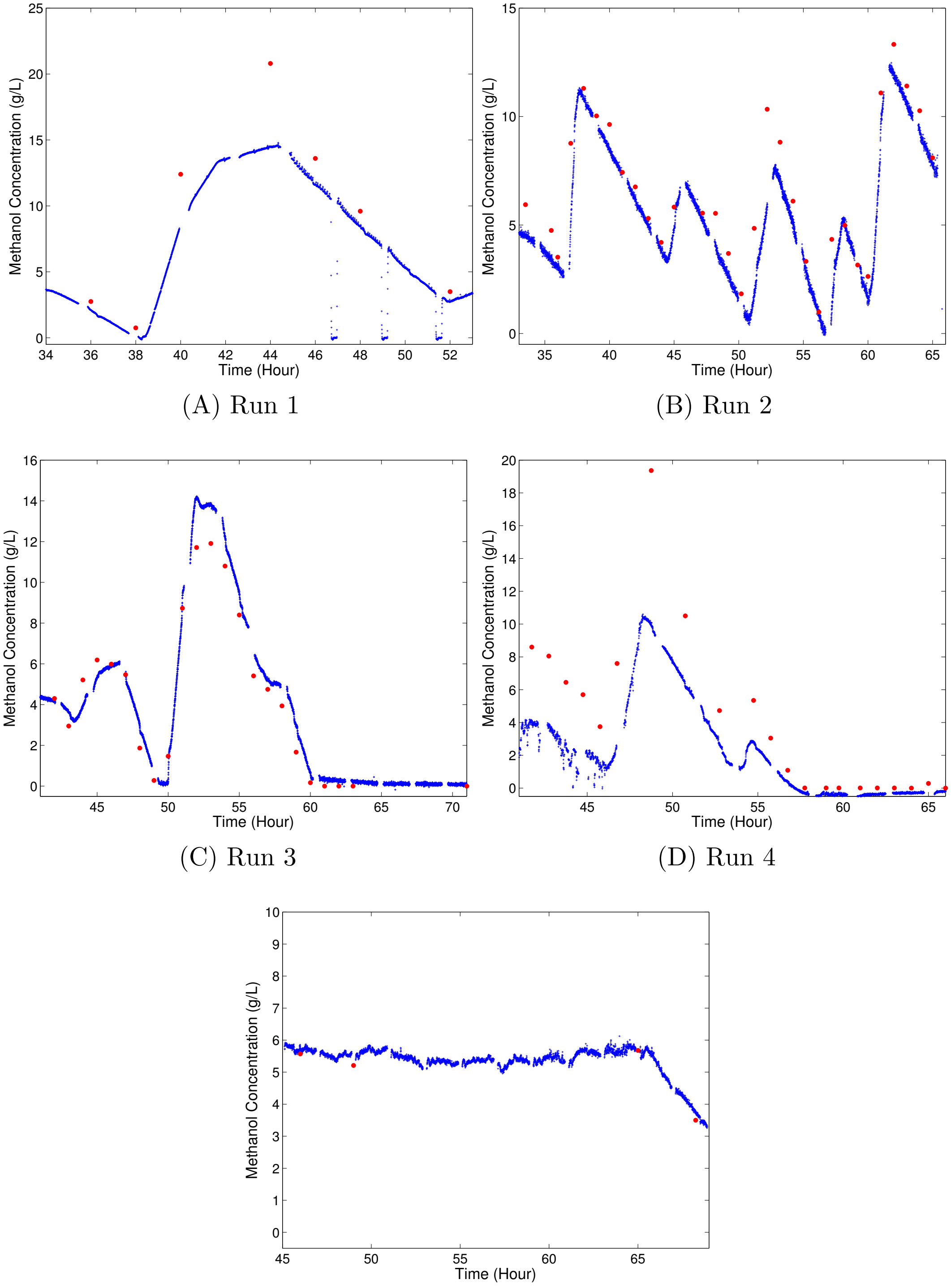
(B) Run 2

(C) Run 3

(D) Run 4

(E) Run 5

Figure 6.4. Prediction results for ACLS models used in glycerol predictions in the bioreactor runs. In each figure, the blue symbols denote the predicted concentrations from the NIR measurements. The red dots are the reference concentrations.



(A) Run 1

(B) Run 2

(C) Run 3

(D) Run 4

(E) Run 5

Figure 6.5. Prediction results for ACLS models used in methanol predictions in the bioreactor runs. In each figure, the blue symbols denote the predicted concentrations from the NIR measurements. The red dots are the reference concentrations.

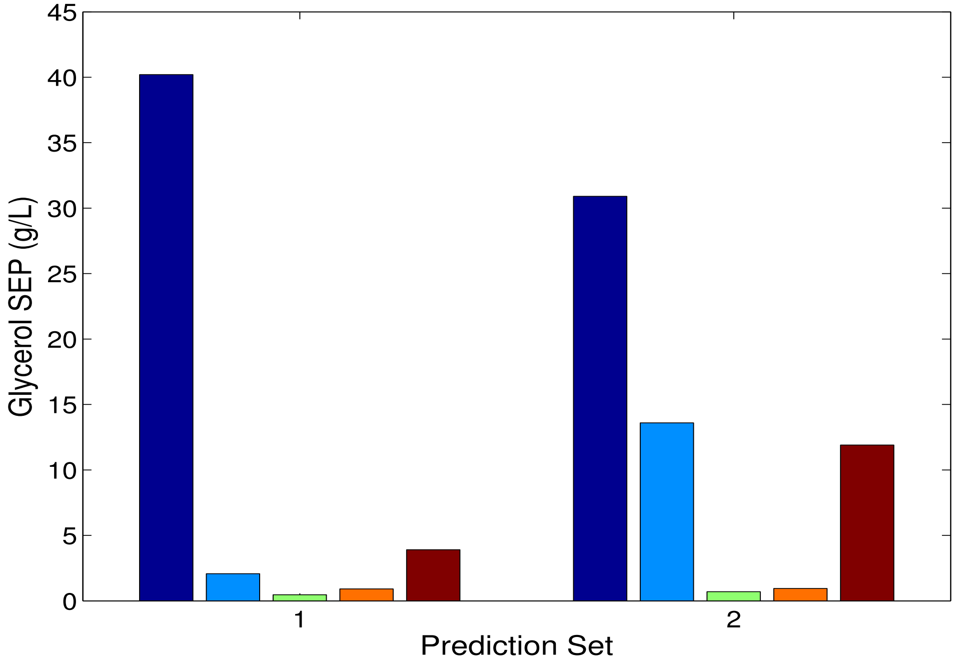
changed from 100 to 600 cm^{-1} with an increment of 20 cm^{-1}. The selection of the number of latent variables for the PLS model used an F-test as before. The maximum number of latent variables was 15. A leave-10%-out cross-validation was employed.

For the laboratory data, the grid search was performed on the calibration set. The bioreactor data used the synthetic monitoring spectra to seek the best ACLS model for each run. The obtained SEP values are plotted in Figures 6.6 and 6.7. Without digital filtering, the ACLS did not perform well with the laboratory data and in the glycerol predictions in the bioreactor data, but it provided a competitive result for methanol monitoring. The PLS algorithm did not produce a stable model for glycerol but worked acceptably for methanol.

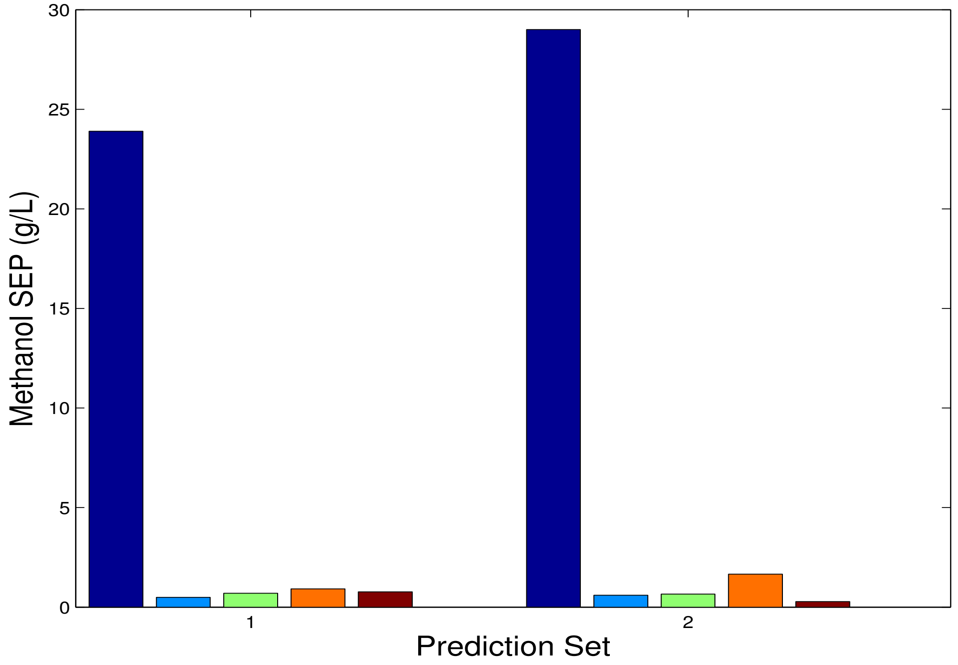
6.5 Conclusions

In this study, we established a protocol for the process monitoring of batch fermentation of *Pichia pastoris* cells. The purpose of the monitoring was to control the carbon source of either cell or protein growth. An in-line method is more desired for the reason of automatic sampling, less labor, and real-time continuous response. The tubing system enables continuous data collection with automatic sampling. The filter-based AOTF spectrometer provides a useful measurement platform for this application because of its rugged and compact properties.

In data analysis, the conventional PLS regression method requires a sufficient number of reference samples to build a calibration model, as well as a chemical system that is stable over time. Therefore, PLS modeling is not readily applicable to this

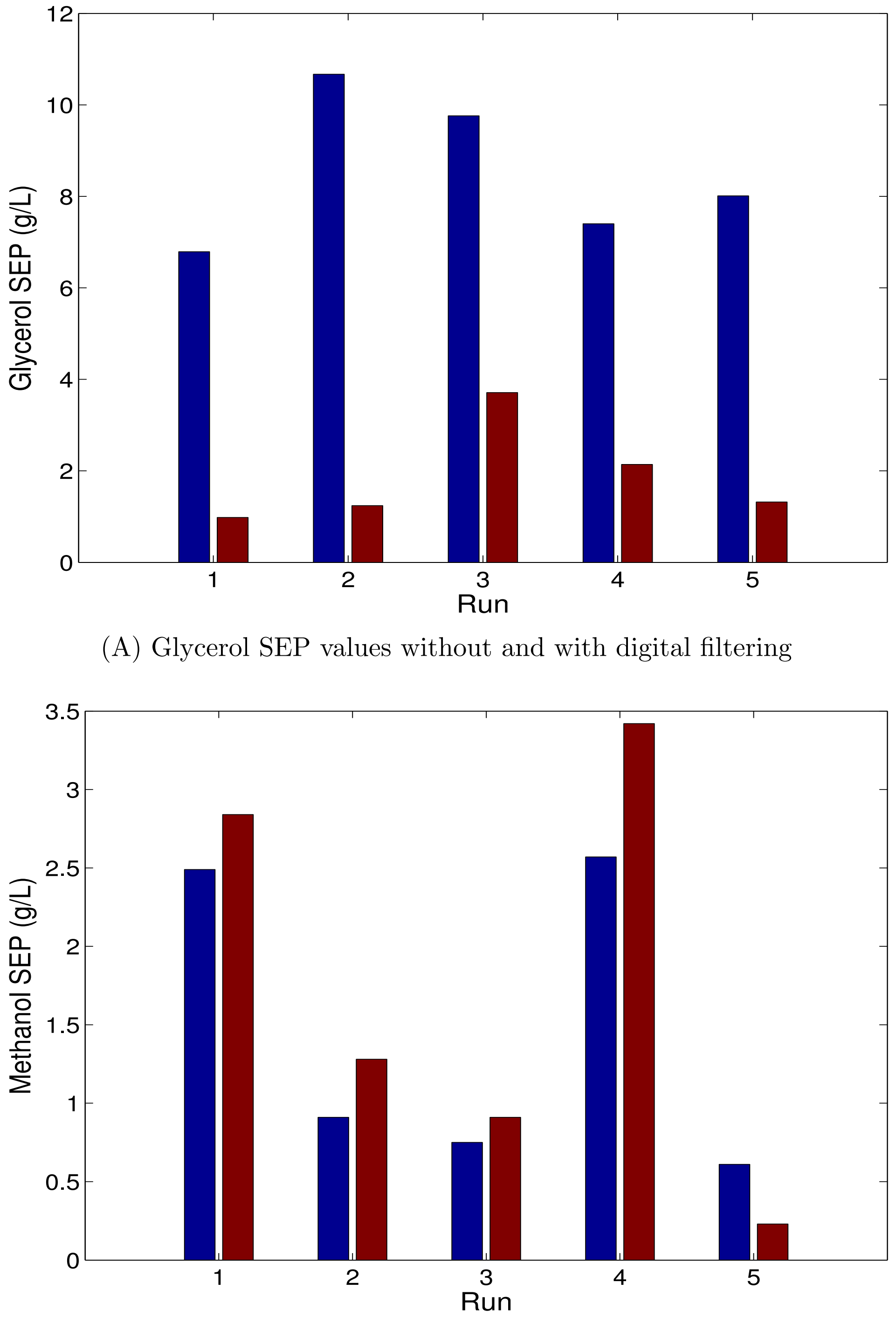


(A) Glycerol SEP values without and with digital filtering



(B) Methanol SEP values without and with digital filtering

Figure 6.6. Compared prediction results (SEP) without (blue) and with (green for $ACLS_1$, orange for $ACLS_2$) digital filtering for the laboratory data. The PLS results are shown in cyan (with digital filtering) and red (without digital filtering). (A) Glycerol results; (B) Methanol results. The $ACLS_1$ results with filtering provide the most consistent performance overall.



(A) Glycerol SEP values without and with digital filtering

(B) Methanol SEP values without and with digital filtering

Figure 6.7. Compared prediction results (SEP) for ACLS models without (blue) and with (red) digital filtering for the bioreactor data. (A) Glycerol results; (B) Methanol results. The results with filtering provide the most consistent performance overall.

bioreactor monitoring application due to the small number of reference measurements available and the dynamic nature of the sample matrix both within and between runs. In this work, an optimized digital filtering/ACLS method was evaluated for use in performing continuous monitoring. The ACLS approach uses the pure-component spectrum of the analyte of interest and other representative component spectra either known or extracted from the background matrix to build the total component matrix. The constructed matrix is used for further prediction. The ACLS regression only needs the pure-component spectra and spectra of the background matrix for calibration, thereby avoiding the need for large quantities of reference data. The modeling approach is also consistent with updating the model at the start of each bioreactor run.

Signal processing, specifically Chebyshev Type II bandpass filtering, was found to be useful as a spectral preprocessing tool. To obtain the best filter design, PSO was used to optimize the filter design variables along with the spectral range submitted to the ACLS calculation. During the optimization process, reference measurements are required to evaluate each possible filter and spectral range, and the optimization has to be completed at an early stage of the process in order that the model be available for use in predicting unknown concentrations. Few, if any, reference points can be assumed to be available other than the known starting composition of the base media. For this reason, this research employed a synthetic data generation step to obtain a set of monitoring spectra for use in driving the model optimization. The approach proved workable and the combination of filtering and ACLS produced good prediction results for both the laboratory and bioreactor data.

One issue that could not be successfully addressed in this work is the need for updating the model as the bioreactor process evolves. For example, digital filters optimized at the beginning of the bioreactor run may no longer be optimal as the run progresses. Updating of the parameters may be needed to re-establish an optimal calibration model. Various methods were attempted to implement this concept but a well-performing solution could not be found. This remains an important area of study for future work.

CHAPTER 7
AUTOMATED CALIBRATION PROTOCOL FOR CONTINUOUS BIOREACTOR MONITORING BASED ON WAVELET PREPROCESSING

7.1 Introduction

In Chapter 6, we have discussed the importance and application of the batch process. Continuous monitoring can be extremely valuable in helping to ensure a successful process application in the pharmaceutical or manufacturing industries. Among potential monitoring tools, spectroscopic methods provide direct chemical information and are nondestructive, thereby enabling in-line process monitoring. Near infrared analysis is especially attractive because of its compatibility with aqueous solutions and easy sample preparation. Therefore, NIR spectroscopy is well-suited for use in implementing automated process monitoring.

A protocol for continuous monitoring via NIR spectroscopy has been established in the previous chapter. To monitor a process by NIR measurements, the classical way is to obtain a group of calibration data points (spectra and corresponding reference concentrations for the target analyte(s)) , build a PLS calibration model and then apply that model to future data for as long as prediction performance is maintained. An update or recalibration of the model is then performed and the cycle repeats.[21]

However, for a real process, the concentration changes might be slow, inducing a long and tedious task in data collection. Furthermore, for limited reference data collections, the obtained concentration range might be not wide enough to interpolate to a representative concentration profile.

To address these limitations, the research described in Chapter 6 employed ACLS models that take advantage of known pure-component spectra and are less reliant on the collection of large numbers of calibration samples and associated reference measurements. The performance of the ACLS models was improved by application of a preprocessing bandpass digital filter that was tuned to the experimental conditions of the current run. A data synthesis procedure was employed to generate a set of monitoring samples for use in optimizing both the filter bandpass and the spectral range submitted to the ACLS model.

The choice of bandpass digital filtering as a spectral preprocessing tool is based on a significant literature that supports its use with NIR data.[100,117,118] It is useful in the removal of potential interference signals from the background matrix and can help to minimize spectral variation associated with changes in sample temperature. In Chapters 5 and 6, we have applied bandpass digital filtering to both laboratory and bioreactor data. Other signal processing methods, for example derivative analysis[119,120] and wavelet analysis, can also play an effective role in the suppression of unwanted spectral features. Wavelet analysis has been widely applied in various spectroscopy fields, including nuclear magnetic resonance (NMR), ultraviolet-visible (UV-vis), and Raman.[121–124] In applications in infrared spectroscopy, the wavelet transform is useful in removing fluctuating backgrounds and in improving classification and quantification accuracy.[59,125,126]

The wavelet transform uses a wavelet function to decompose an input signal into a set of wavelet coefficients termed approximations and details. The approximations can be further decomposed into a new set of approximations and details. Depending on the decomposition level, the wavelet coefficients of the approximation or details have different frequency and time resolution than the original signal. The multi-resolution property of wavelets is helpful in the analysis of the signal, for example background correction, noise removal, and peak separation. With a certain wavelet function and proper selection of the decomposed wavelet coefficients, the analyte spectral features can be extracted, thereby inducing a better quantitative result. Details about the principles of the wavelet transform have been introduced in Chapter 3.

In this chapter, we continue the investigation of the real-time continuous monitoring of the bioreactor runs of *Pichia pastoris* cells. Glycerol and methanol are the carbon sources and the analytes of interest in the study and are used for cell growth and protein production, respectively. Here, the wavelet transform is used as the spectral preprocessing tool in place of digital filtering. Both the laboratory simulated data and the bioreactor process data were studied. Optimization is still a necessity in order to seek the proper wavelet function and the corresponding decomposed coefficients. The ACLS modeling technique was again used to obtain estimated glycerol and methanol concentrations. The utility of the wavelet transform in improving prediction performance is compared to that of digital filtering.

7.2 Experimental

The work described in this chapter employed the same AOTF spectrometer described previously in Chapters 5 and 6. The collection of laboratory and bioreactor data was described in Chapter 6. In the data analysis employing wavelet preprocessing, both of the non-cell laboratory data and the bioreactor data were studied.

The data partitioning used was also the same as described in Chapter 6. The laboratory data were divided into calibration (Day 1) and prediction sets (Days 16 & 17). The calibration data were also used as the monitoring spectra in the optimization runs. For the bioreactor data, the problem of limited reference points persisted in this study. The synthetic data generation strategy based on using mathematically generated monitoring spectra for use in driving the optimization proved to be useful in the work described in Chapter 6 and was also employed here.

All computational work was performed under MATLAB (version 7.4, The MathWorks, Inc., Natick, MA) running on a Dell Precision 670 workstation (Dell Computer Corp., Austin, TX) operating under Red Hat Enterprise Linux WS (Red Hat, Inc., Raleigh, NC). The wavelet transform and PSO calculations were implemented with the Matlab wavelet toolbox (version 6.7, The MathWorks, Inc.) and the public-domain Particle Swarm Optimization Toolbox developed by Brian Birge (version 2.5) and available through the file exchange maintained by The MathWorks, Inc.

7.3 Data Analysis Strategy

7.3.1 Overview of Data Analysis

With the air reference as the background, each collected spectrum was converted into absorbance units for use in further analysis. The point spacing was reduced from 0.67 cm^{-1} to 4 cm^{-1}. Data preprocessing focused on the use of the discrete wavelet transform (DWT). The DWT uses a wavelet function to decompose an input signal into hierarchical sets of approximations and details. The signal can then be reconstructed using some subset of the details in order to remove unwanted components. Preprocessed spectra were then used in the construction of quantitative ACLS models for glycerol and methanol. Models based on PLS were also constructed with the laboratory data for comparison. Optimization using the PSO method was again employed in finding optimal values for the signal processing and model generation parameters.

7.3.2 Implementation of Optimization with the Wavelet Transform

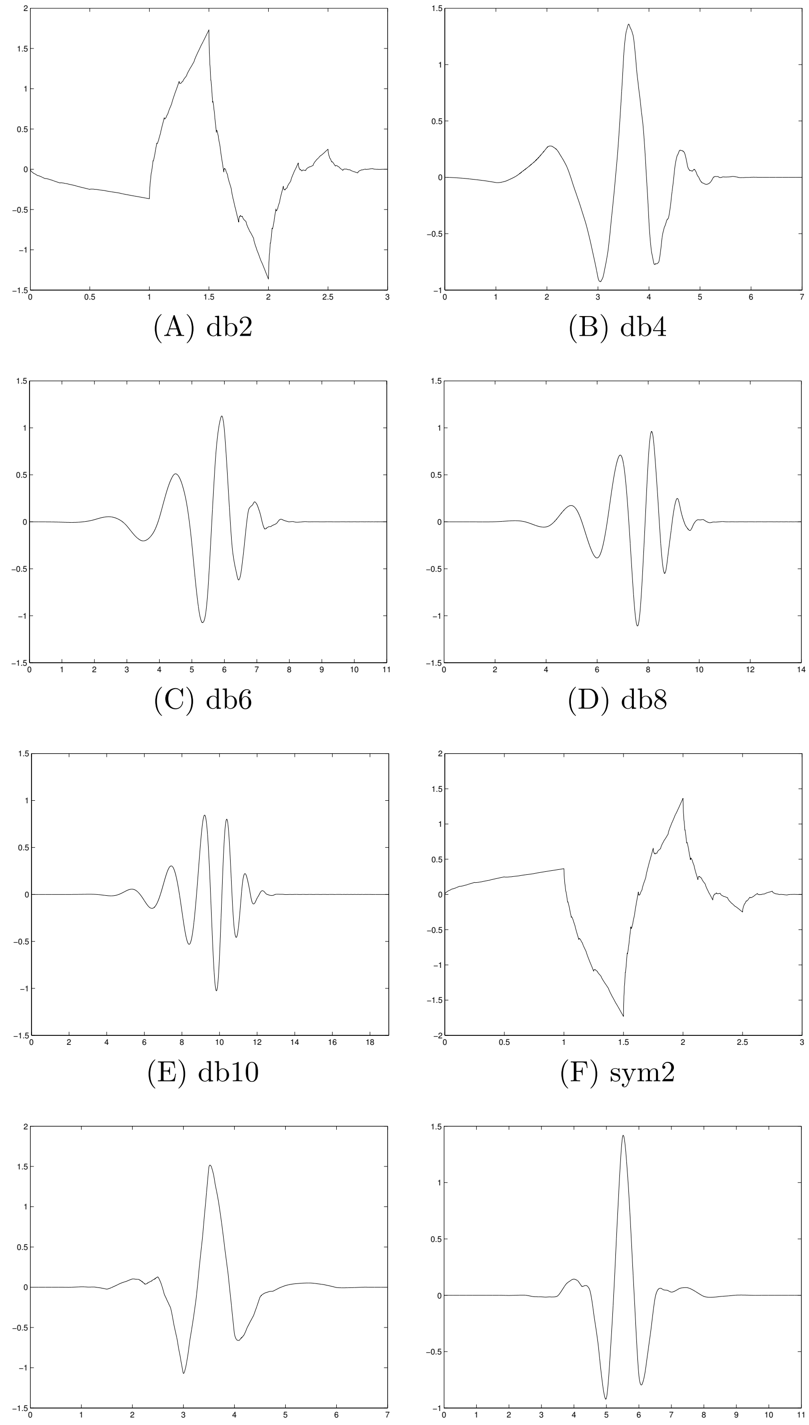
The most frequently used wavelet functions include the Daubechies, Symmlet and Biorthogonal families. The family specification is also termed the mother wavelet. Individual functions within each family are designated by an additional order parameter. Functions from different families have different properties in terms of shape, symmetry, support and vanishing moment.[57,58].

As noted above, during the wavelet transform process, the signal is decomposed into one approximation and one detail component. The obtained approximation can then be further decomposed into new approximation and detail components. This process continues for a specified number of levels of decomposition, n. This leads to one final approximation component and n detail components. Taken together, the approximation and n details can be used to reconstruct the original input signal (spectrum) exactly. If one or more of these components are not used in the reconstruction, an altered spectrum is obtained in which some of the original information has been suppressed.

The approximation coefficients obtained from the final level of decomposition are assumed to carry the broad background information that underlies the narrower spectral features that are superimposed on the background. In the reconstruction process used here, the approximation was discarded. Detail coefficients from different levels were then selected by the optimization procedure In the wavelet transform of this study, wavelet functions from two wavelet families, 'Daubechies' (db) and 'Symmlet' (sym), were used. Specifically, db2, db4, db6, db8, sym2, sym4, sym6, and sym8 functions were involved in the optimization. The number after the family specifier is the order parameter that identifies the specific wavelet function used. Figure 7.1 plots the wavelet functions used in this study. The maximum decomposition level was 6 with the minimum at 2. In the PSO calculations, the selection of the reconstructed coefficients was performed in a binary mode. Each level of details coefficients was assigned either '1' or '0' initially, where '1' indicated that the corresponding level was included and '0' meant that the level was omitted from the reconstruction of the spectrum. In the optimization, the total number of optimized parameters for the signal processing was eight, including the wavelet function, the level of decomposition and the selection parameters for each level. For decomposition levels lower than 6, the PSO still generated six binary numbers for reconstruction, but the last several numbers were idle. For example, a 4-level decomposition would use the first four binary numbers in the reconstruction and ignore the last two.

After the signal processing, a quantitative analysis was performed to predict glycerol and methanol concentrations. The spectral range is an essential factor for both ACLS and PLS models. For the PLS algorithm used with the laboratory data, the number of latent variables included in the model is also a key parameter. The optimization of these regression methods has been introduced previously in Chapters 5 and 6 and the same procedures were used here. A maximum of 15 latent variables was allowed in the PLS models and the F-test procedure described in Chapters 5 and 6 was used in the evaluation of each set of parameters to determine the optimal model size.

The optimized parameters of the wavelet transform and multivariate regression are listed in Table 7.1. For one group of ACLS models developed for the laboratory data in Chapter 6, the soy peptone, yeast nitrogen base (YNB), and yeast extract components were included in the optimization and allowed to be chosen. This procedure was also used in the work reported here.



(A) db2

(B) db4

(C) db6

(D) db8

(E) db10

(F) sym2

(G) sym4

(H) sym8

Figure 7.1. Wavelet functions used in optimization. The notation “dbN” represents functions from the Daubechies family, while “symN” specifies functions derived from the Symlet mother wavelet.

Table 7.1. Optimized parameters in wavelet transform, augmented CLS and PLS

Method	Parameters
Wavelet Transform	Functions: db2[a], db4, db6, db8, sym2[b], sym4, sym6, sym8 Decomposition level: 2 - 6 Level of details selected for reconstruction: 2 - 6
Augmented CLS	Spectral range: 4800 to 4200 cm^{-1} in steps of 5 cm^{-1} Additional components: soy peptone, yeast extract, YNB
PLS	Spectral range: 4800 to 4200 cm^{-1} in steps of 5 cm^{-1} Number of latent variables: 5 to 15

[a] Daubechies family with order 2. Other Daubechies functions are specified analogously

[b] Symlet family with order 2. Other Symlet functions are specified analogously.

7.4 Results and Discussion

Examples of spectra before and after the wavelet transform are shown in Figure 7.2. Table 7.2 lists the wavelet function used, the corresponding decomposition level and the selected levels of details employed in the reconstruction, and the spectral range used in the regression step. In the table, ACLS$_1$ denotes the algorithm based on augmented CLS in which the background was represented by the collected matrix spectrum at the beginning of the experiment. Then, the $\mathbf{K_a}$ matrix was composed of the pure-component spectra of glycerol, methanol, sorbitol, and the matrix spectrum. In the ACLS$_2$ method, additional pure components (soy peptone, yeast extract, and YNB) were involved in the optimization such that the model was given the option of including these terms. Therefore, the corresponding $\mathbf{K_a}$ would contain more components.

The SEP values describing the prediction performance of each method are listed in Table 7.3. The corresponding correlation plots of calibration and prediction are shown in Figures 7.3 and 7.4. In the correlation plots, the left column shows the predicted values versus the reference values from the calibration in blue plus signs. On the right are shown the correlation plots for the two prediction sets, in which the blue plus signs represent the first set and red symbols indicate the second.

From the error values and the correlation plots, the three methods all calibrate very well for the glycerol models. Compared to the PLS method, however, the ACLS_1 and ACLS_2 methods provide much better prediction results and a more stable correlation between the estimated and reference values. The PLS prediction has the issue of bias in the prediction results. This suggests that the calibration data are not globally representative such that the derived PLS loadings adequately carry forward to the prediction data collected more than two weeks later.

In the methanol predictions, although the PLS method provided a much better calibration, the prediction performance is not as good as that of the calibration. The systematic bias problem observed with the glycerol model persisted. Whereas the ACLS_2 method could predict better than PLS in terms of the SEP values, slight bias in the prediction can still be observed.

Generally speaking, for both glycerol and methanol, the ACLS model with component selection included in the optimization provided good performance with the calibration and prediction data. The PLS method could calibrate very well but the computed model could not extend to the prediction data collected subsequently. It appears an update of the PLS model is required to make it compatible with the

Table 7.2. Wavelet analysis and regression models for non-cell laboratory data

(a) Glycerol

Model	Wavelet Transform Analysis			Regression Model	
	Wavelet Function	Decomposition Level	Levels of Details in Reconstruction	Spectral Range (cm^{-1})	Note
$ACLS_1$	db4	4	[1,2]	4595 - 4260	
$ACLS_2$	db4	6	[1,5]	4735 - 4200	Peptone
PLS	db4	6	[1,2,3,5,6]	4935 - 4395	LV[a]= 9

(b) Methanol

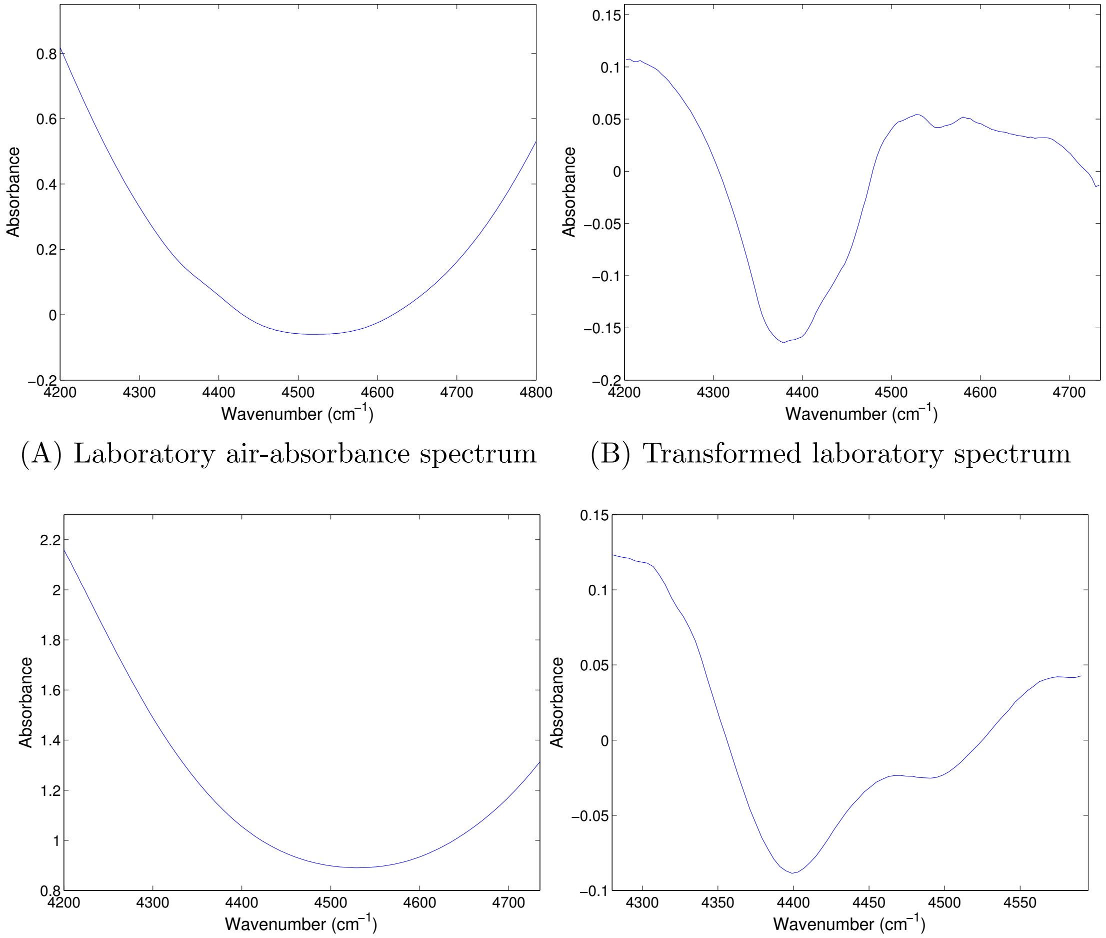
Model	Wavelet Transform Analysis			Regression Model	
	Wavelet Function	Decomposition Level	Levels of Details in Reconstruction	Spectral Range (cm^{-1})	Note
$ACLS_1$	db8	2	[1,2]	4750 - 4390	
$ACLS_2$	db10	3	[2,3]	4800 - 4440	none
PLS	db6	6	[1,3,4,5]	4725 - 4365	LV[a]= 7

[a] Number of latent variables in PLS model.

prediction data. This illustrates the principal problem with the use of the PLS approach for a continuous monitoring application in which little time is available for the collection of extensive amounts of calibration data. By contrast, the ACLS can easily incorporate information from the prediction data through the collection of a very simple set of solutions (pure-component solutions and a sample of the base media) on the prediction day. This data collection is considered feasible as part of a start-up protocol for a bioreactor run.

7.4.1 Process Monitoring Results for the Bioreactor Data

The optimization results obtained for the bioreactor data are listed in Table 7.4 for glycerol results and Table 7.5 for methanol results. Both tables list the optimized wavelet function parameters and spectral ranges used with the ACLS algorithm. As in



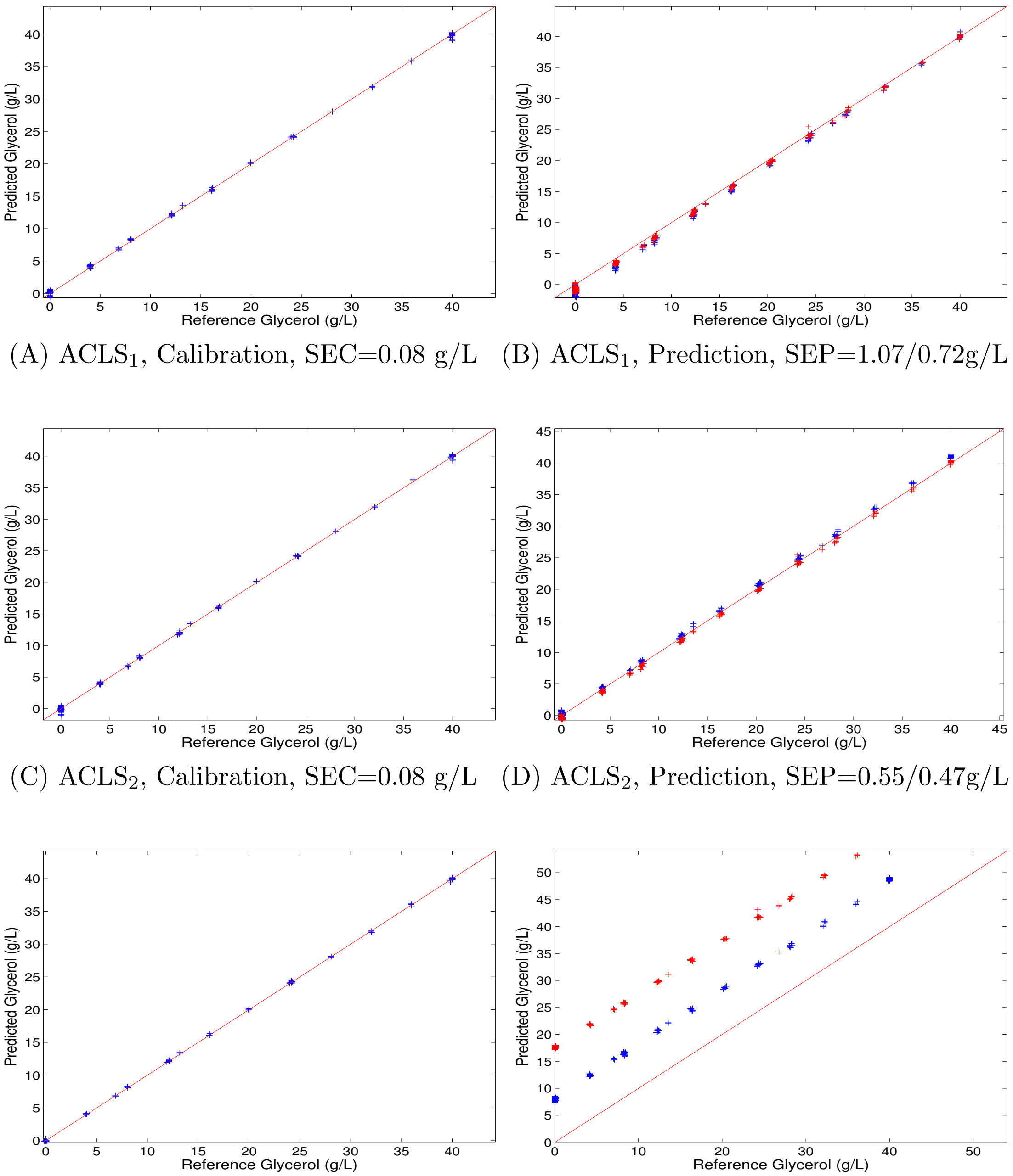
(A) Laboratory air-absorbance spectrum

(B) Transformed laboratory spectrum

(C) Bioreactor air-absorbance spectrum

(D) Transformed bioreactor spectrum

Figure 7.2. Spectra before and after wavelet transform with the optimized parameters. (A) Laboratory spectrum with air as background. (B) Reconstructed spectrum after wavelet transform. The transform was performed based on the following optimized parameters: wavelet function = db4, decomposition level = 6, reconstruction levels = (1,6). (C) Air absorbance spectrum from bioreactor Run 1. (D) Reconstructed spectrum after wavelet transform with function db10, 6 decomposition levels, and reconstruction based on details from levels (1,2,5,6).

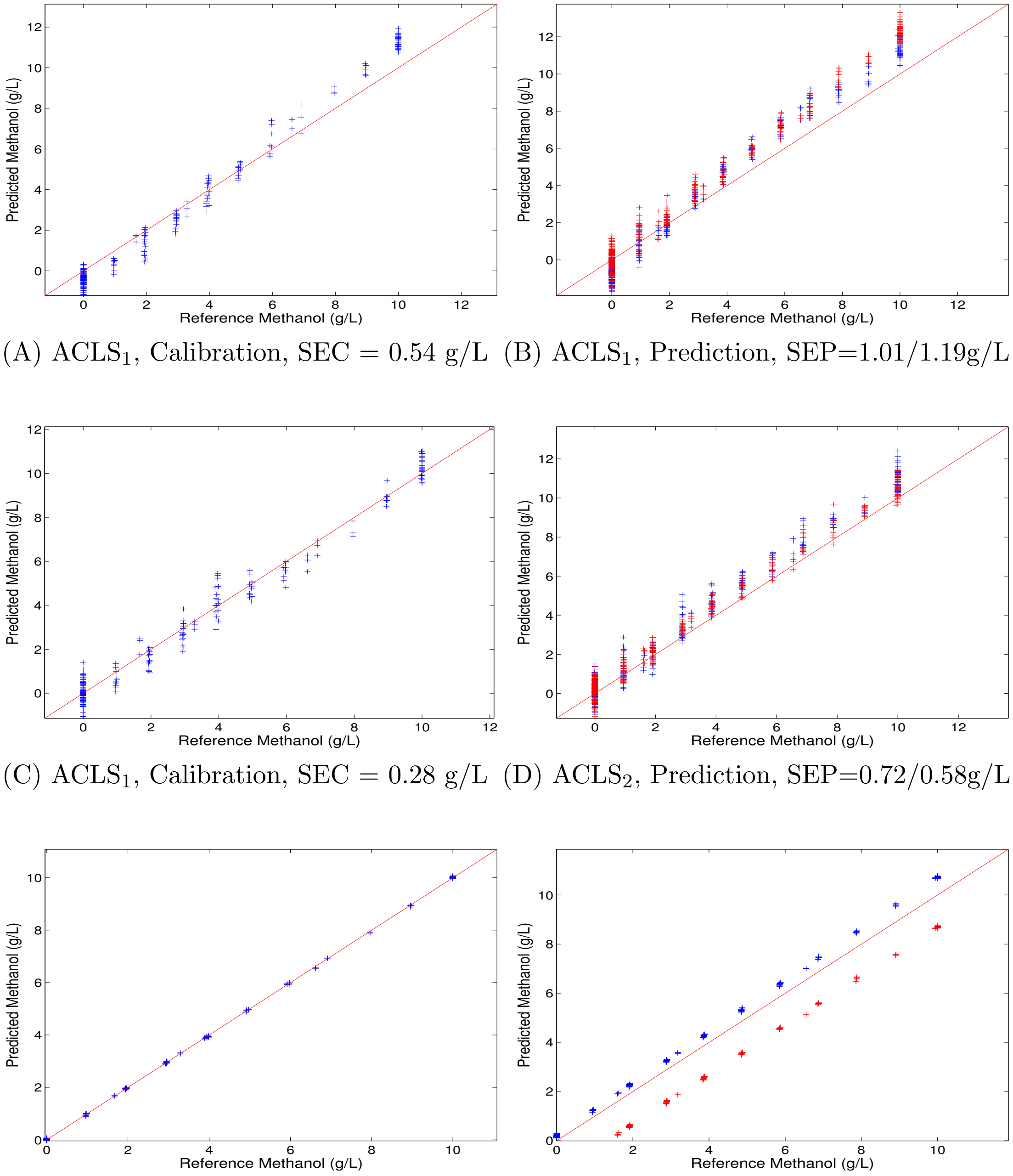


(A) ACLS$_1$, Calibration, SEC=0.08 g/L (B) ACLS$_1$, Prediction, SEP=1.07/0.72g/L

(C) ACLS$_2$, Calibration, SEC=0.08 g/L (D) ACLS$_2$, Prediction, SEP=0.55/0.47g/L

(E) PLS, Calibration, SEC = 0.08 g/L (F) PLS, Prediction, SEP = 8.23/17.5 g/L

Figure 7.3. Prediction results for glycerol in laboratory data with different algorithms. The subfigure caption shows the corresponding algorithm, data set and error value. ACLS$_1$ is the augmented CLS algorithm. This algorithm used the three major components (glycerol, methanol and sorbitol) and the spectrum of the background matrix in assembling the $\mathbf{K_a}$ matrix. The ACLS$_2$ method also involved a selection of additional components in optimization from soy peptone, YNB, and yeast extract. In the prediction plots, the blue signs denote the results for prediction set 1 and the red symbols denote the second prediction set. Values of SEP are given for both prediction sets.



(A) ACLS$_1$, Calibration, SEC = 0.54 g/L (B) ACLS$_1$, Prediction, SEP=1.01/1.19g/L

(C) ACLS$_1$, Calibration, SEC = 0.28 g/L (D) ACLS$_2$, Prediction, SEP=0.72/0.58g/L

(E) PLS, Calibration, SEC = 0.03 g/L (F) PLS, Prediction, SEP = 0.40/1.36 g/L

Figure 7.4. Correlation plots for methanol predictions with different algorithms for laboratory data. The subfigure caption shows the corresponding algorithm, data set and error value. ACLS$_1$ is the augmented CLS algorithm. This algorithm used the three major components (glycerol, methanol and sorbitol) and the spectrum of the background matrix in assembling the $\mathbf{K_a}$ matrix. The ACLS$_2$ method also involved a selection of additional components in optimization from soy peptone, YNB, and yeast extract. In the prediction plot, the blue symbols indicate the results for prediction set 1 and the red symbols denote the second prediction set.

Table 7.3. SEP values in g/L for non-cell data

(a) Glycerol

Data Set	$ACLS_1$	$ACLS_2$	PLS
Calibration	0.08	0.08	0.08
Prediction 1	1.07	0.55	8.23
Prediction 2	0.72	0.47	17.5

(b) Methanol

Data Set	$ACLS_1$	$ACLS_2$	PLS
Calibration	0.54	0.28	0.03
Prediction 1	1.01	0.72	0.40
Prediction 2	1.19	0.58	1.36

Chapter 6, no PLS models were attempted because of the lack of sufficient calibration data and the dynamic nature of the bioreactor data both within and between runs. The synthetic data used for monitoring the optimization of the ACLS model was deemed not accurate enough for use in generating PLS models.

In using the ACLS method, the $\mathbf{K_a}$ of the glycerol model included pure-component spectra of glycerol and the spectra (i.e., the first two principal components) extracted from the background matrix during the first five hours of the run. During the methanol feed, the glycerol was still possibly present. Thus, the $\mathbf{K_a}$ matrix for methanol has both methanol and glycerol pure-component spectra, as well as the two components derived from the background spectra taken during the five hours before the methanol feed. As in Chapter 6, an updated $\mathbf{K_a}$ matrix was used for each bioreactor run.

Figure 7.5 depicts the glycerol monitoring results for the five runs. In each sub-figure, the blue plus signs represent the predicted concentrations based on the spectroscopic method. The red dots denote the values measured by the reference method. The plot covers from the beginning of the runs until the glycerol feed ends. Due to the suspected inaccuracy of the reference method at high concentrations, the prediction results at the initial part of the run exhibit a shift from the reference values, especially for Run 3.

In order to quantify the prediction results, the standard error value of the prediction was calculated by matching the time when spectra and reference points were collected. As estimated, it usually took 10 minutes for the solution to flow from the fermentor to the spectrometer. Hence, a 10-minute lag is included in calculating the SEP. Because the spectra collected during the first five hours were used in generating the monitoring data and performing the optimization for glycerol, the data points collected in this time slot were excluded in calculating the glycerol SEP values. The SEP values are listed in Table 7.6.

The prediction results for methanol are presented analogously in Figure 7.6. The plots show the results during the period after the methanol feed started until the last reference point was collected. The calculation of the SEP values still considered a 10-minute delay. The results of this calculation are shown in Table 7.6.

Observing the correlation plots and SEP values, an effective job was done overall tracking the glycerol and methanol concentrations. Run 3 was problematic for glycerol monitoring after hour 20, with the spectroscopic results being positively offset from the reference values. In addition, some of the runs have an initial reference glycerol concentration different from the assumed 14 g/L.

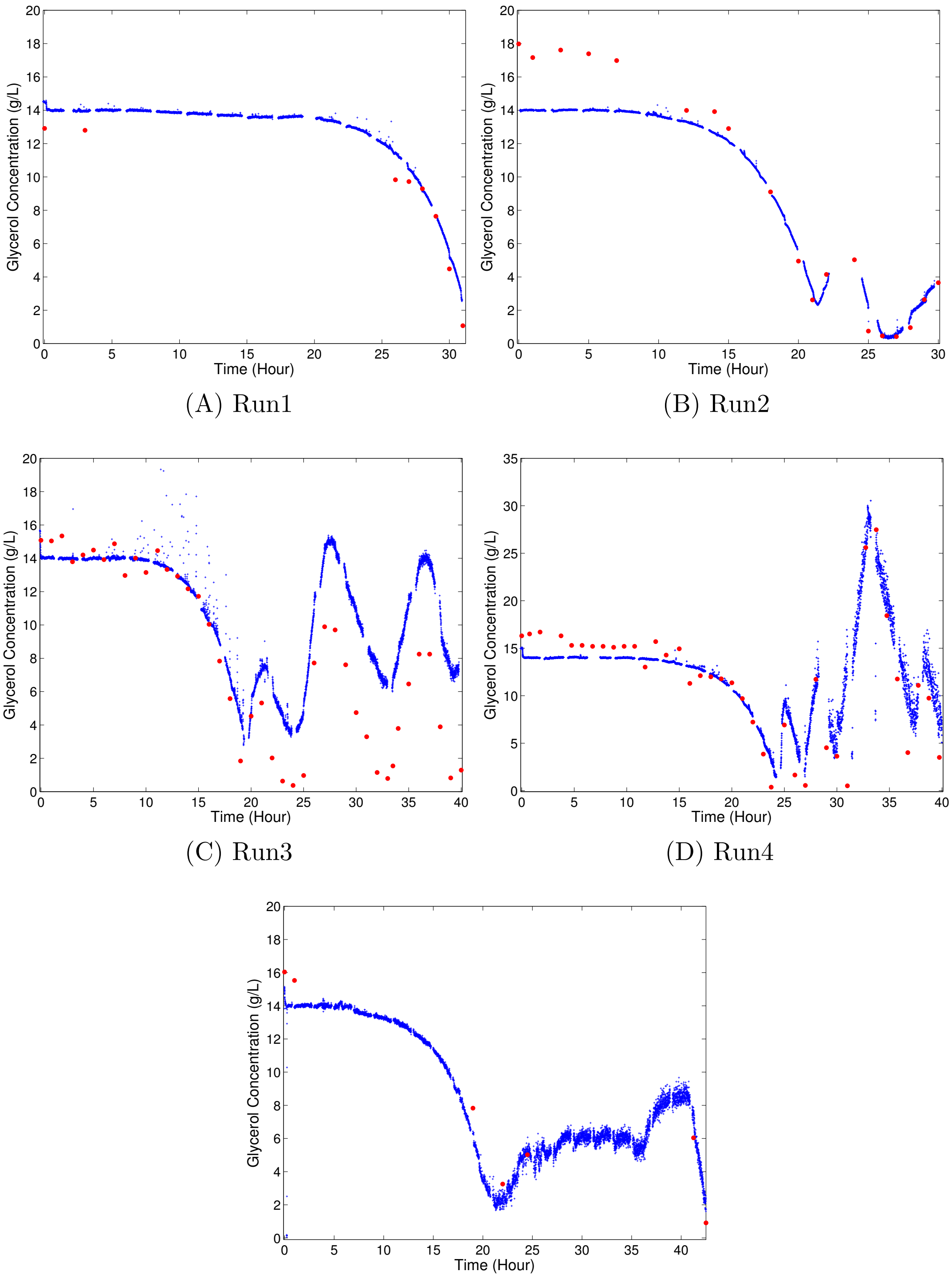
Table 7.4. Wavelet analysis and regression model parameters used in glycerol prediction for the bioreactor data

Run	Wavelet Transform Analysis			CLS
	Wavelet Function	Decomposition Level	Levels of Details in Reconstruction	Spectral Range (cm^{-1})
1	db10	6	(1,2,5,6)	4595 - 4280
2	db4	6	(2,3,5,6)	4640 - 4260
3	db6	4	(3,4)	4790 - 4365
4	db4	4	(3)	4650 - 4325
5	db8	6	(1,4,6)	4475 - 4335

Table 7.5. Wavelet analysis and regression model parameters used in methanol prediction for the bioreactor data

Run	Wavelet Analysis			CLS
	Wavelet Function	Decomposition Level	Levels of Details in Reconstruction	Spectral Range (cm^{-1})
1	db6	6	(5,6)	4630 - 4315
2	db10	5	(5)	4750 - 4310
3	db4	6	(1,2,4,5,6)	4705 - 4320
4	db8	5	(2,4,5)	4660 - 4205
5	db8	6	(1,3,4,5,6)	4555 - 4320

With regard to the methanol results, except Run 1 and Run 4, the SEP values were all lower than 1 g/L for the ACLS models. The larger SEP in Run 1 is caused by the large variation from the reference value at around 44 hours. The result for Run 4 is similar to the results obtained from Chapter 6. For this run, the predicted results from the NIR measurements are biased, all lower than the reference values until hour 58.



(A) Run1

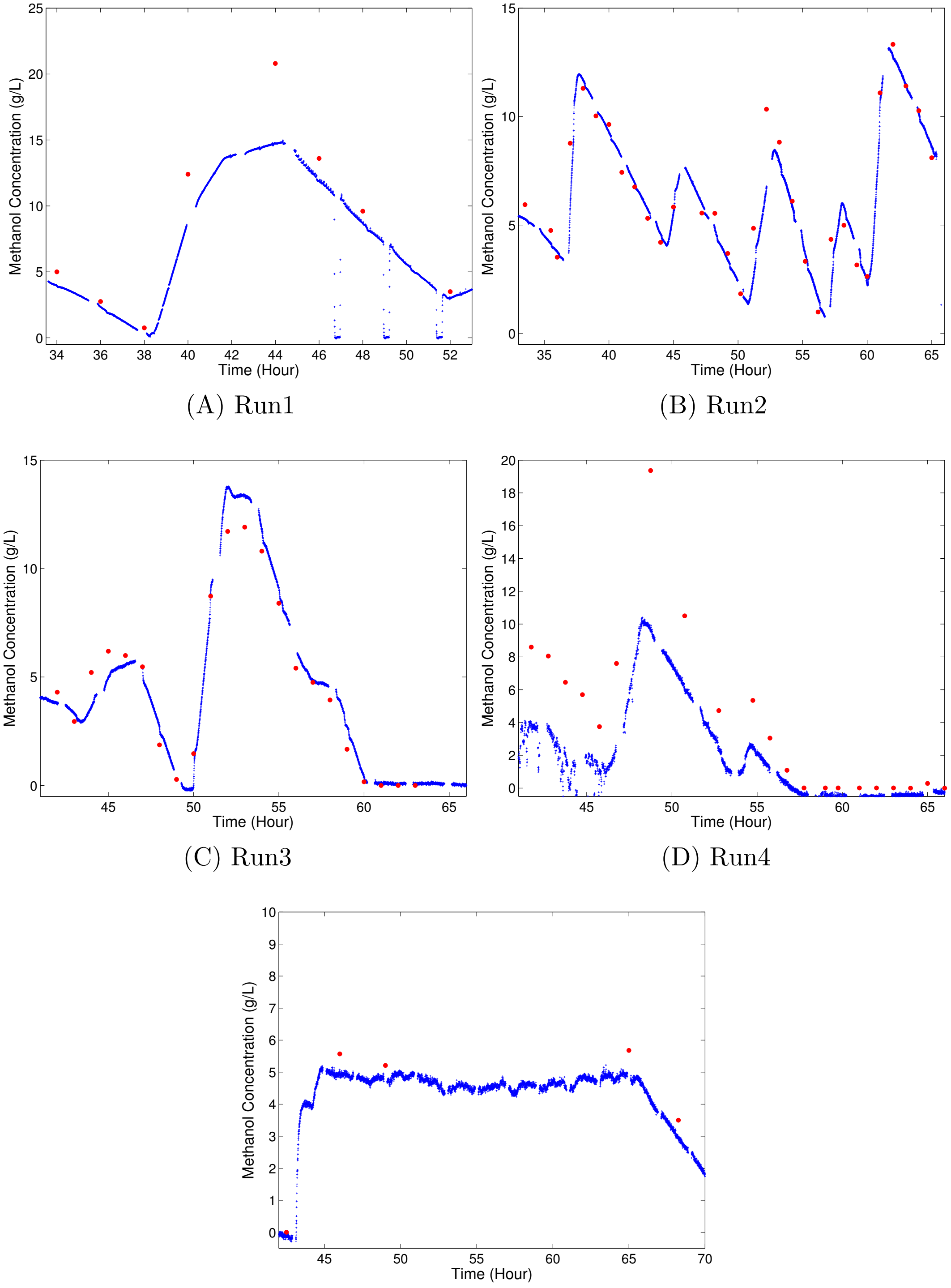
(B) Run2

(C) Run3

(D) Run4

(E) Run5

Figure 7.5. Results for ACLS models in glycerol predictions for the bioreactor runs. In each figure, the blue signs are the predicted concentrations from the NIR measurement and the red dots indicate the reference concentrations.



(A) Run1

(B) Run2

(C) Run3

(D) Run4

(E) Run5

Figure 7.6. Results for ACLS models in methanol predictions for the bioreactor runs. In each figure, the blue signs are the predicted concentrations from the NIR measurement and the red dots indicate the reference concentrations

Table 7.6. Prediction results for bioreactor monitoring

Run	Glycerol SEP (g/L)	Methanol SEP (g/L)
1	0.84	2.63
2	1.70	0.75
3	2.90	0.79
4	2.21	3.42
5	1.11	0.65

7.4.2 Comparison with Digital Filtering Methods

7.4.2.1 Results

In Chapter 6, we employed Chebyshev Type II digital filtering as the signal processing tool. The results of non-cell laboratory data monitoring are shown in Figure 6.2 and 6.3 and the SEP values are listed in Table 6.2. In this chapter, the wavelet transform method replaced digital filtering. The comparison results with regard to the SEP values are shown as bar plots in Figures 7.7A and 7.7B for the laboratory data. Figures 7.7C and 7.7D remove the PLS results to provide easier comparisons between the results obtained with the ACLS models.

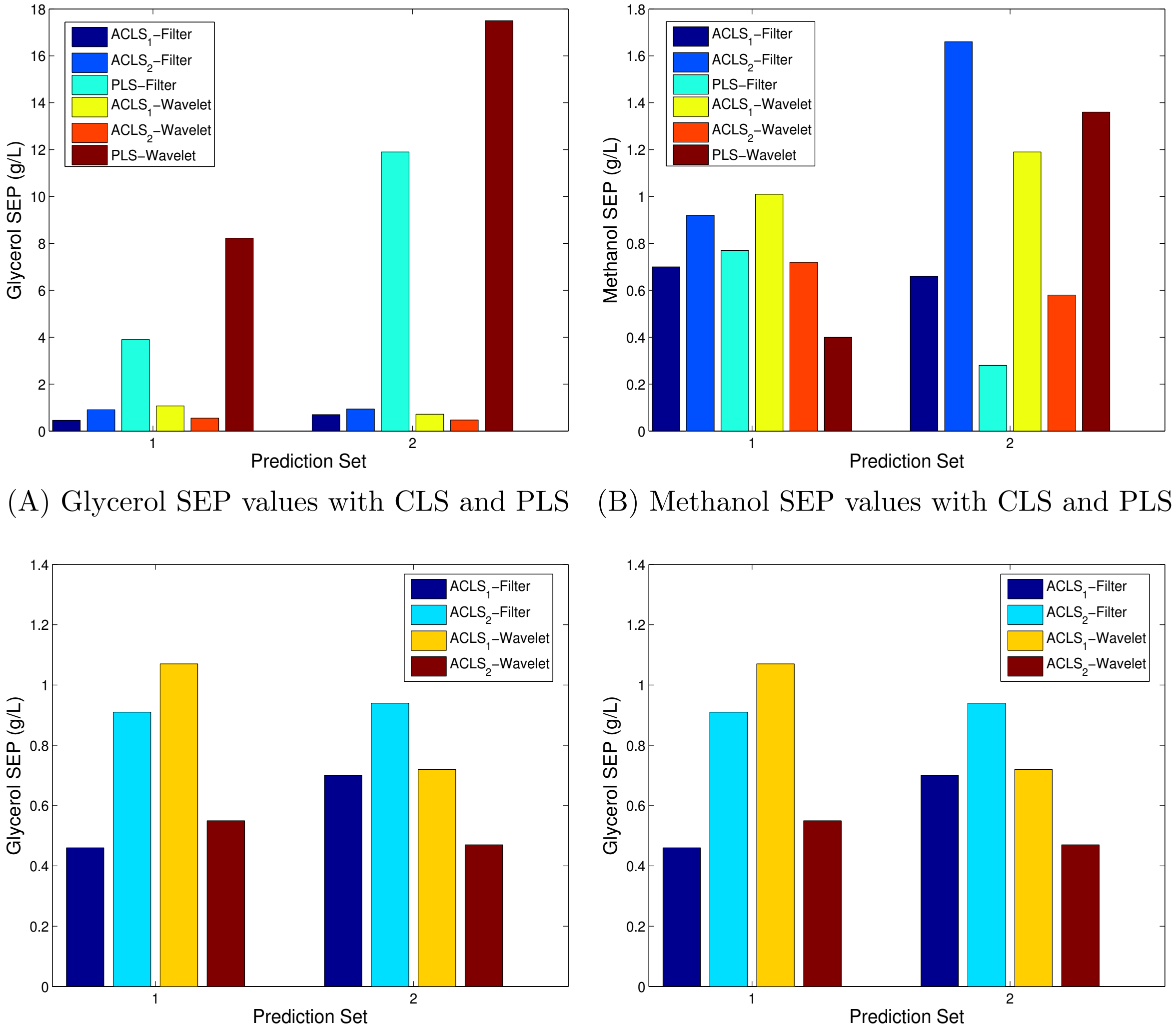
Obviously, the PLS model could not provide a good prediction for glycerol two weeks after the collection of the calibration data. Similarly, the methanol prediction is not stable for the second set. Regarding the ACLS methods, the ACLS$_1$ method with digital filtering and the ACLS$_2$ model with the wavelet transform provided comparable prediction performance for both glycerol and methanol. The ACLS$_2$ model with wavelet preprocessing predicted slightly better in general.

To monitor the bioreactor process, we established a process monitoring protocol. The two signal processing tools were tested under the same protocol. The prediction results and the corresponding SEP values via the digital filtering method are shown in Figure 6.4, Figure 6.5 and Table 6.5. Figures 7.8A and 7.8B show the SEP values of these two methods for use in predicting glycerol and methanol concentrations, respectively.

With the ACLS method, the two signal processing methods provided similar results for the bioreactor data. Digital filtering worked better for Runs 2 and 4, while the wavelet method predicted better in Runs 1, 3 and 5. With respect to the methanol predictions, the two signal processing tools provided similar results for each of the runs. The issue of systematic offset in Run 4 persisted. Generally speaking, the wavelet method worked better than the digital filtering procedure except in Run 5.

7.4.2.2 Computational Time Considerations

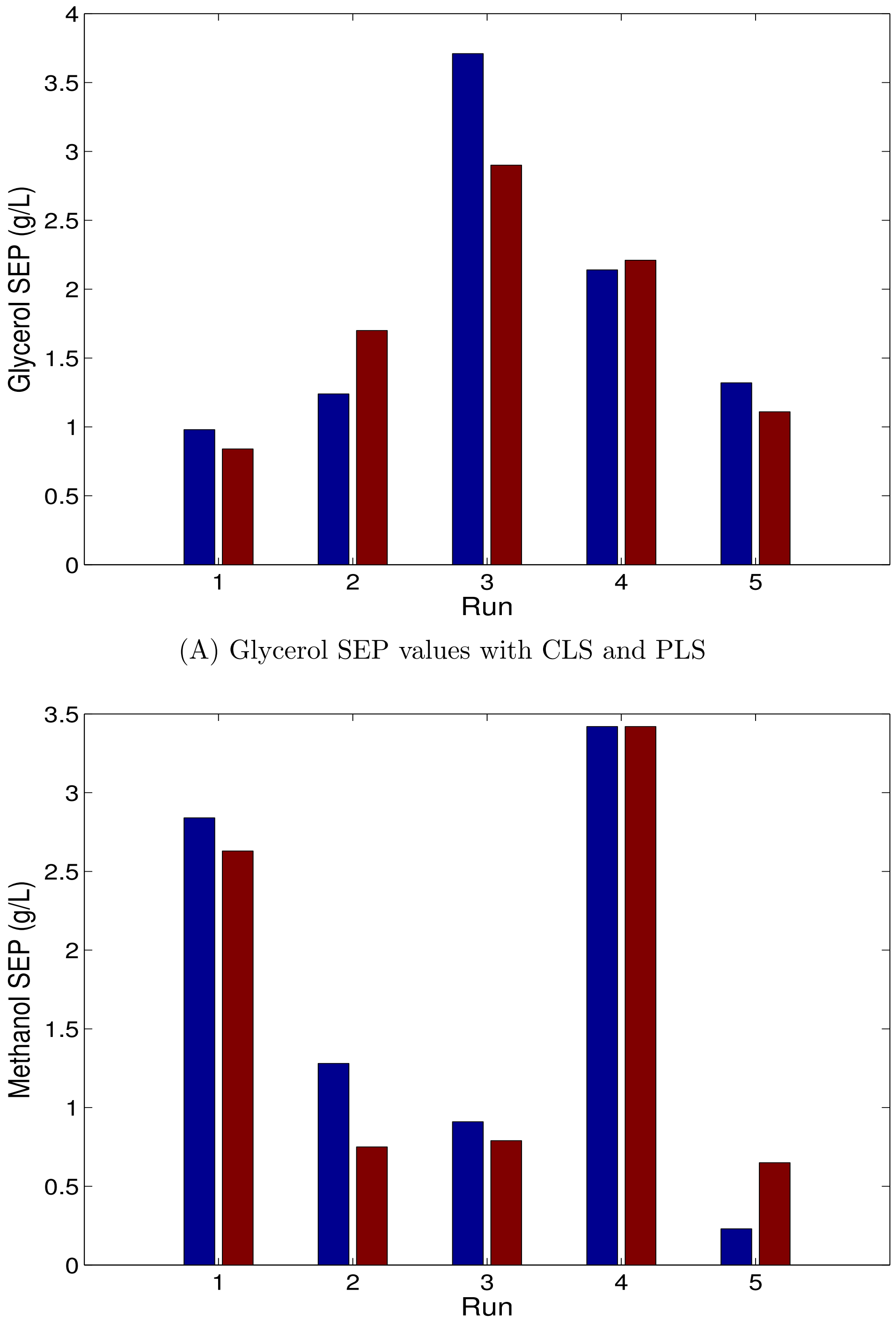
In this chapter and Chapter 6, PSO was employed in selecting the best model for prediction. To apply this method in a real application, the optimization time has to be taken into consideration. The optimization process is divided into two steps: data processing and optimization. Factors that could affect the calculation efficiency in the data processing step include the spectral matrix size, time cost in signal processing and regression. The optimization can be affected by the population size, the particle dimensionality, the search space of the particles, and the total number of iterations taken for the model to converge to the optima. Because the optimization has a random component, different initial populations could induce different total numbers



(A) Glycerol SEP values with CLS and PLS (B) Methanol SEP values with CLS and PLS

(C) Glycerol SEP values with CLS (D) Methanol SEP values with CLS

Figure 7.7. Compared prediction performances for laboratory data. Panels (A) and (B) compare the SEP values for glycerol and methanol, respectively. They show the results of the various combinations of signal processing and modeling algorithms. Panels (C) and (D) compare the two signal processing methods with only the ACLS methods. The ACLS$_2$ method with wavelet processing provides the best overall prediction results.



(A) Glycerol SEP values with CLS and PLS

(B) Methanol SEP values with CLS and PLS

Figure 7.8. Compared prediction performances with respect to the signal processing methods for glycerol (A) and methanol (B) in the bioreactor data. The digital filtering method is shown in blue bars and the wavelet transform results are in red. The wavelet processing outperforms filtering in three of the five runs.

of iterations. To compare the respective computational times, each method was tested with a fixed number of iterations (e.g., 50), without consideration of whether or not the optimization converged. All calculations were based on a 200 × 150 spectral matrix (i.e., 200 spectra containing 150 points) with 50 initial particles in the PSO calculation.

Optimization of the wavelet/ACLS parameters was approximately 38 times slower than the corresponding optimization of the digital filtering/ACLS parameters. Two factors determine this difference. First, the digital filtering process is much faster than the wavelet decomposition/reconstruction step. In addition, the wavelet transform method has more parameters (8) to optimize than the digital filtering approach (6). Overall, wavelet processing outperformed digital filtering in terms of prediction performance. However, a fast computational platform would be required to implement this method in practice.

7.4.3 Comparison Results for Methods without Signal Processing

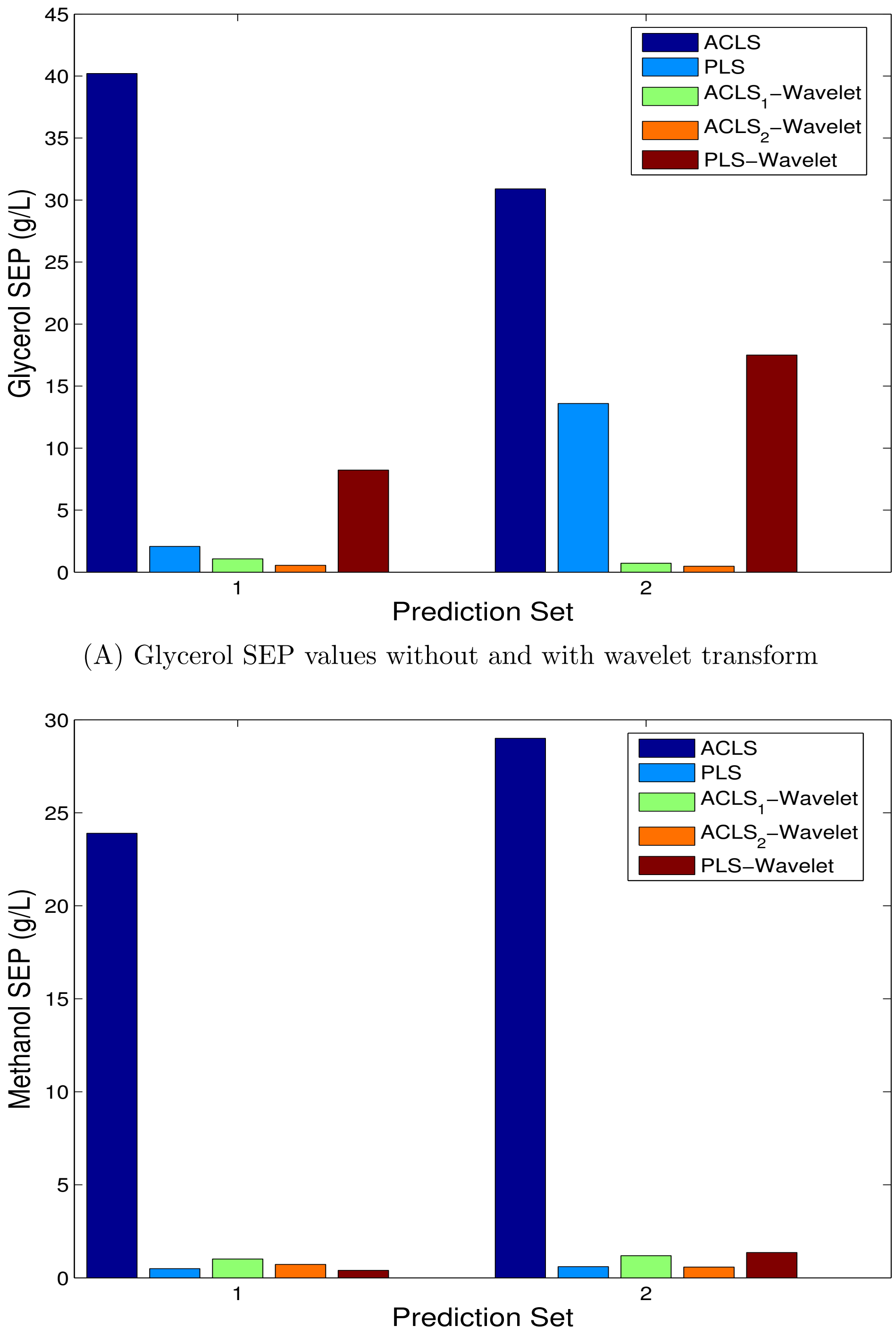
Results obtained from this work are also compared with models built without signal processing. Similar to Chapter 6, the ACLS method without signal processing involved only a spectral range selection step. This was done by use of the calibration spectra of the laboratory data and the synthetic monitoring spectra from the bioreactor data. The spectral ranges were optimized by a grid search method. In the grid search, the spectral range was selected over 4800 to 4200 cm^{-1}. The spectral range varied from 200 to 600 cm^{-1} with a 20 cm^{-1} increment. The moving step of the range was 5 cm^{-1}. The $\mathbf{K_a}$ matrices are the same as those used with the signal processing method. No component optimization was performed for the laboratory data. The SEP values obtained without the use of signal processing are plotted in Figures 7.9 and 7.10.

The models built without signal processing could provide similar results to those obtained with the wavelet method for only a few cases, notably methanol prediction in the bioreactor process. However, without the wavelet transform, stable predictions could not be obtained in general. The signal processing step is clearly critical in helping to remove components from the data that cannot be adequately represented in the $\mathbf{K_a}$ matrix used in obtaining the predicted concentrations.

7.5 Conclusions

In this chapter, we investigated further the spectroscopic process monitoring of bioreactor runs of *Pichia pastoris*. By controlling the concentrations of the glycerol and methanol feedstocks, cell growth and protein production can be enhanced. A successful continuous monitoring protocol of methanol and glycerol concentrations can enable automated control of the glycerol and methanol feeds. The dynamic nature of the bioreactor process, both within a run and between runs, makes the calibration component of the NIR measurement extremely challenging, however.

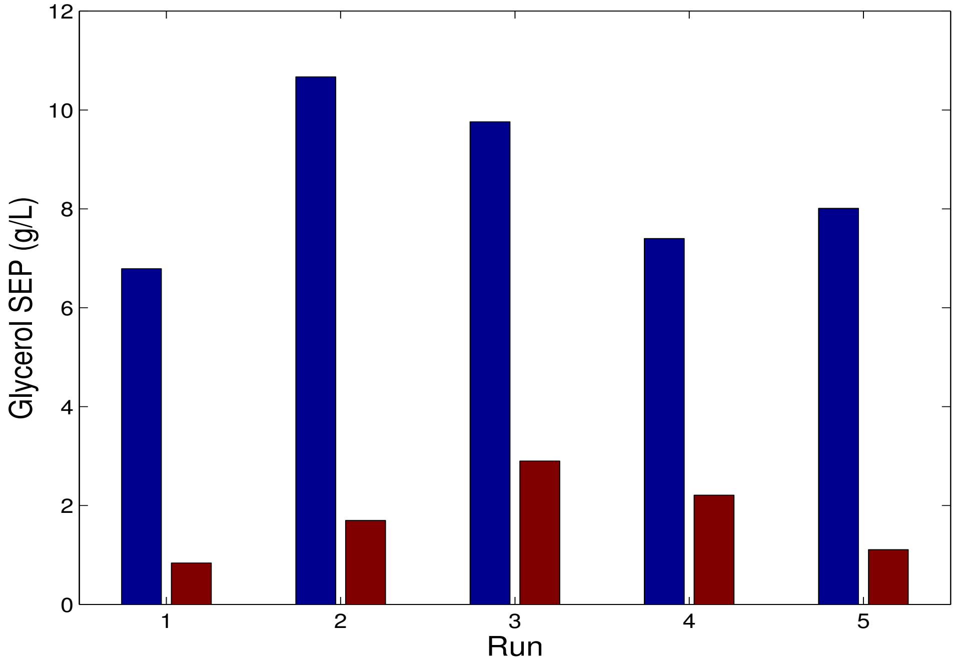
To help stabilize the calibration, the wavelet transform was employed as a spectral preprocessing technique in this study. Optimization is still required to seek the appropriate wavelet parameters to extract useful spectral features. With the



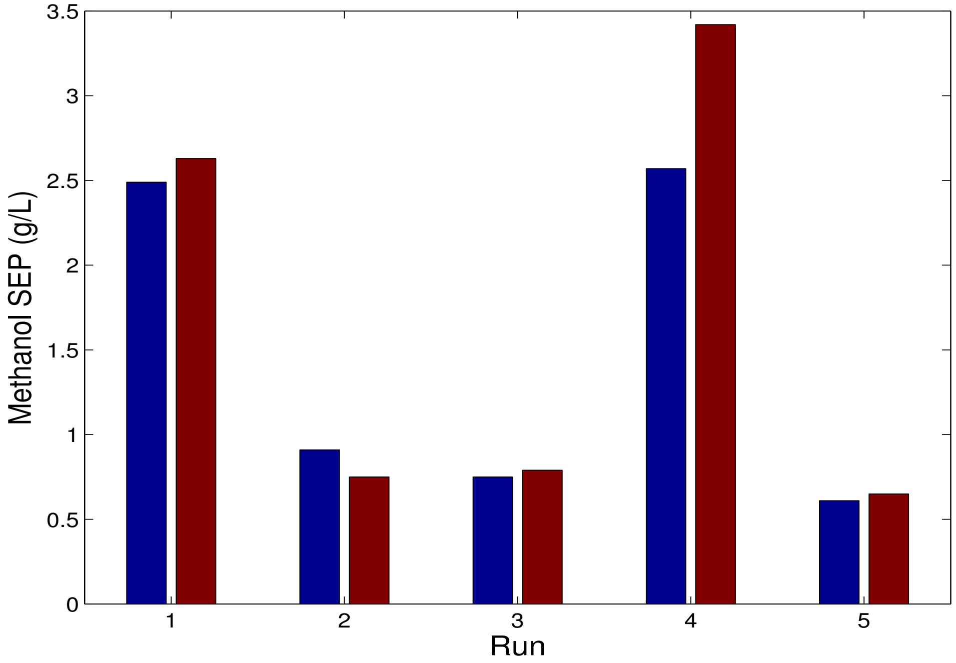
(A) Glycerol SEP values without and with wavelet transform

(B) Methanol SEP values without and with wavelet transform

Figure 7.9. SEP values for predictions without (ACLS in blue, PLS in cyan) and with (ACLS in green and orange, PLS in red) the wavelet transform for the laboratory data. The wavelet preprocessing step provides greater benefit to the ACLS models than the PLS models. This suggests the wavelet processing is effective in removing components from the spectra that have not been explicitly encoded in the $\mathbf{K_a}$ matrix.



(A) Glycerol SEP values without and with wavelet transform



(B) Methanol SEP values without and with wavelet transform

Figure 7.10. Bar plots of comparison SEP values for ACLS prediction results without (blue) and with (red) the wavelet transform for the bioreactor data. The wavelet preprocessing provides clear benefit for the glycerol models but does not improve the prediction results for methanol.

successful strategy of spectral synthesis of monitoring data introduced in Chapter 6, ACLS models could be optimized separately for each prediction day in the laboratory data or each bioreactor run. Similar to the digital filtering methodology employed in Chapter 6, the wavelet transform was able to suppress components of the data that could not be adequately encoded in the ACLS model. Models built without wavelet preprocessing were not as stable in prediction.

Regarding the choice of preprocessing method, digital filtering (Chapter 6) and the wavelet transform (Chapter 7) gave somewhat similar results, with the wavelet preprocessing being slightly better overall. The increased computational time associated with wavelet processing vs. digital filtering and the larger search space required for the wavelet parameters made the wavelet/ACLS models more computationally intensive than those based on digital filtering/ACLS. This is a potentially important consideration when choosing a computational platform for use with the method.

Further work could be performed to speed the optimization of the wavelet parameters. For example, the initial population size is the primary factor to affect the load of the calculation. The use of fewer particles could dramatically save time during each iteration. This may result in the need for more iterations, however, and could decrease the likelihood of finding the global optimum. Another strategy would be decrease the matrix size of the spectra fed into the optimization. Possible changes could include larger point spacing, fewer spectra in the background matrix, or a smaller sample size in the monitoring spectra.

CHAPTER 8
CONCLUSIONS AND FUTURE WORK

This dissertation focused on the investigation of quantitative analysis methods for use with infrared spectroscopy in two challenging applications: passive remote sensing and process monitoring. Both of these applications involve dynamic measurement environments with unstable spectral backgrounds. Furthermore, the collection of large quantities of calibration data for use in building quantitative models is both expensive and impractical. Through the development of methods in signal processing, data synthesis, and optimization, workable quantitative models were obtained in both applications.

In remote sensing, quantitative work is made difficult by the high cost and effort required to produce known releases of chemicals for use in acquiring calibration spectra. In addition, background spectral measurements are made difficult due to the radiance changes caused by spectrometer movement or fluctuations in environmental conditions. This background variation invalidates the traditional computation of analyte-specific absorbance by taking the ratio of analyte and background signals. Such variation in background radiance also limits the quality of any collected calibration data, which is required to be representative of future data to which a computed calibration model will be applied.

In Chapter 4, a simulation method was developed to avoid the required data collection for use in building quantitative models. The synthetic strategy used a radiance model to compute spectra after obtaining the key parameters that described the current experimental scenario. In this study, the temperatures of the analyte and background were the primary factors affecting the signal intensity besides the analyte concentration. This method avoided the collection of background spectra because the calibration model was built solely on the basis of the synthetic single-beam spectra. The only required data collection was one stable sample release for use in estimating the sample temperature and the spectra of a blackbody source at two temperatures for use in estimating two parameters related to the spectrometer. Without a requirement for the collection of experimental data, the calibration data set could be generated quickly through the synthesis procedure.

Results from both laboratory and field data showed that when the temperature difference between the analyte plume and the background was higher, predicted concentrations were more accurate because of the higher absorption or emission intensity. Overall, reasonable quantitative prediction performance was obtained in an application in which the measured signals are inherently unstable. While the field application described here employed a ground-based spectrometer looking upward at the effluent from an emission stack, it is projected that the methodology could work similarly in an implementation in which a released chemical plume is detected in a downward-looking mode from above.

Besides the remote sensing study, data analysis strategies in NIR spectroscopy were also implemented for use in process monitoring. The broad and overlapping features found in NIR spectra make quantitative analysis challenging because spectral selectivity is limited. This lack of selectivity necessitates the use of multivariate calibration models. Reliance on such models presents practical challenges in industrial applications because of the need for the models to perform well without the need for frequent or tedious recalibration. In the latter part of this thesis, the objective was to develop calibration protocols for dynamic systems, using both a laboratory-based flowing system and a real process bioreactor system. Signal processing techniques were employed, facilitated by numerical optimization methods to search for parameter combinations that optimized the calibration protocols.

In Chapter 5, a calibration method was developed that combined digital filtering, partial-least squares (PLS) regression, and particle swarm optimization (PSO) for use in continuous monitoring. Instead of a conventional FT-IR spectrometer, an AOTF-based filter spectrometer was employed due to its compatibility with an industrial environment. Both long- and short- term data sets were investigated.

In the data analysis, buffer and air spectra were compared for use as spectral backgrounds. The air spectra were actual internal reference spectra collected in tandem with each sample spectrum, while the buffer spectra were collected in a block before the sample data collection started. The air background spectra provided more stable predictions than the buffer spectra for the long-term study, while no significant difference was observed for the short-term data.

Signal processing was investigated in this chapter to remove non-analyte spectral information. The Chebyshev Type II bandpass filter was applied in the digital filtering preprocessing. A study using principal component analysis (PCA) illustrated the effectiveness of the filter in removing time-based variation from the spectra. In the quantitative analysis step, an F-test was used to avoid overfitting the PLS model, a common problem with this modeling method. A key component of the methodology was the use of PSO to combine the optimization of the digital filter and PLS modeling parameters.

An extension of this work was the use of a model updating method to improve calibration stability in the long-term study. By adding buffer spectra into the original calibration data, a new calibration could be built through the same optimization process on each prediction day. Improvement in prediction performance was observed in lower bias between the reference and predicted concentration values. Future research aspects of this work might include the study of a different digital filter design or preprocessing method. In this study, there were four components in each sample besides the background buffer. A more complicated solution system would be helpful to test the robustness and stability of this data analysis method.

In Chapters 6 and 7, the research moved forward to a real process bioreactor system. In the monitoring of the bioreactor process, only limited reference points could be obtained, leading to difficulties in building PLS calibration models. To address this limitation, the augmented classical least-squares (ACLS) method was explored to build calibrations. Using the CLS method requires knowledge of the constituents of the chemical system under study and either their corresponding pure-component spectra or representative composite spectra of several components. This method enables the construction of a calibration model without a large collection of calibration spectra and associated reference measurements.

In these chapters, signal processing methods were still implemented. In addition to the digital filtering method, wavelet transforms were also investigated. The purpose of the preprocessing step was to remove unnecessary features from the collected spectra before their submission to the calibration model. In this way, it was hoped that concentration estimates would be more stable and less susceptible to changes in the spectral background associated with the dynamic nature of the bioreactor process. Similar to digital filtering, wavelet preprocessing also has different parameter combinations that affect its performance. These include the family and order of the wavelet function, number of levels of decomposition, and how the reconstruction was performed. Optimization with PSO was employed to identify the best preprocessing parameters.

Laboratory data that contained the same chemical constituents as the actual bioreactor runs were studied with the ACLS method in order to characterize its performance and evaluate the optimization of the two signal processing methods. The availability of a set of calibration spectra and reference concentrations also allowed the PLS method to be used.

There were four combinations of preprocessing (digital filtering and wavelet processing) and regression (PLS and ACLS). For long-term predictions, the PLS model was not stable with either preprocessing technique, while the ACLS model still predicted well two weeks after the calibration was optimized. The key difference in the two methods was that the ACLS model could be updated with pure-component and background spectra collected on the prediction day.

For the bioreactor data, however, the limited reference points were not sufficient to guide the optimization. In this case, a synthetic method was designed to create monitoring data for use in guiding the optimization. By combining pure-component spectra and background spectra collected during a time period in which little change in analyte concentration was expected, a set of monitoring spectra was generated. After optimization, both signal processing methods provided similar results in the individual runs, with wavelet processing slightly outperforming digital filtering. A good consistency with the known reference was obtained in general, although a bias in the predicted concentration values was observed during several time periods.

To improve the methodology in the future, testing must be performed with more bioreactor runs for which reliable reference data are available. This will help to provide a better characterization of the stability and robustness of the calibration procedures. In addition, a model updating method needs to be investigated.

As currently implemented, the ACLS calibration model is not flexible in the composition of the components included. As the bioreactor run proceeds, if a new component is introduced (e.g., the addition of a surfactant to reduce foaming), the calibration needs to incorporate that component into the model. This requires an approach in which a pool of possible components is available and a fitting procedure is used to select which to include in the model. It would be desirable to have this procedure be automated rather than requiring intervention by the user.

Given that it is desired to collect very few, if any, reference samples for offline analysis, spectral diagnostics must be available to identify when the model is underperforming. Examination of the spectral residuals obtained from the fit of the ACLS model to the input spectrum may be a key to this diagnostic.

Computational efficiency is also a concern in the practical implementation of the methodology. For example, while the wavelet preprocessing showed some improvement over digital filtering, optimization of the wavelet parameters was very time-consuming. To use the wavelet method practically, several modifications of the optimization can be done: (1) reduce the initial particle size; (2) shrink the search space; and (3) decrease the size of the spectral matrix. Further work should include a study to attempt to streamline the optimization of the wavelet parameters.

In these chapters, the bioreactor process was monitored with a real-time response. Because of the advantages of the spectroscopic method, the developed calibration protocols could be applied in other fields, for example, during a batch process in a pharmaceutical or manufacturing application. The spectral collection can be automated and can greatly reduce the labor required to perform monitoring in a long continuous process. A control system driven by the NIR method could be a key element in helping to optimize and maintain quality control in any number of industrial processes.

Finally, it should be emphasized that the applications addressed in this dissertation were extremely challenging. Obtaining good quantitative performance in passive remote sensing and bioreactor monitoring is nontrivial. There is little or no automated quantitative analysis methodology currently available in either field. The methodology developed in this dissertation provides a significant advance.

REFERENCES

[1] Kowalski, B. *Chemometrics: Theory and Application*; ACS Symposium Series 52: Washington D.C., 1977.

[2] Sharaf, M.; Illman, D.; Kowalski, B. *Chemometrics*; Wiley-Interscience: New York, 1986.

[3] Brereton, R. *Applied Chemometrics for Scientists*; John Wiley & Sons: Chichester, UK, 2007.

[4] Burgess, C. *Valid Analytical Methods and Procedures*; The Royal Society of Chemistry: Cambridge, UK, 2000.

[5] Griffiths, P.; De Haseth, J. *Fourier Transform Infrared Spectrometry*; John Wiley & Sons: Hoboken, NJ, 2007.

[6] Tran, C. *Analytical Chemistry* **1992**, *64*, 971A.

[7] Rathore, C.; Wright, R. *International Journal of Remote Sensing* **1993**, *14*, 1021–1042.

[8] Piccot, S.; Masemore, S.; Ringler, E.; Srinivasan, S.; Kirchgessner, D.; Herget, W. *Journal of the Air & Waste Management Association* **1994**, *44*, 271–279.

[9] Hart, B.; Griffiths, P. *Environmental Science & Technology* **2000**, *34*, 1337–1345.

[10] Hart, B.; Berry, R.; Griffiths, P. *Environmental Science & Technology* **2000**, *34*, 1346–1351.

[11] Hashmonay, R.; Natschke, D.; Wagoner, K.; Harris, D.; Thompson, E.; Yost, M. *Environmental Science & Technology* **2001**, *35*, 2309–2313.

[12] Carter Jr, R.; Thomas, M.; Marotz, G.; Lane, D.; Hudson, J. *Environmental Science & Technology* **1992**, *26*, 2175–2181.

[13] Wan, B.; Small, G. *Analyst* **2008**, *133*, 1776–1784.

[14] Wan, B.; Small, G. *Analyst* **2010**, *136*, 309–316.

[15] Andersson, M.; Josefson, M.; Langkilde, F.; Wahlund, K. *Journal of Pharmaceutical and Biomedical Analysis* **1999**, *20*, 27–37.

[16] Berntsson, O.; Danielsson, L.; Lagerholm, B.; Folestad, S. *Powder Technology* **2002**, *123*, 185–193.

[17] El-Hagrasy, A.; Drennen III, J. *Journal of Pharmaceutical Sciences* **2006**, *95*, 422–434.

[18] Tosi, S.; Rossi, M.; Tamburini, E.; Vaccari, G.; Amaretti, A.; Matteuzzi, D. *Biotechnology Progress* **2003**, *19*, 1816–1821.

[19] Rohe, T.; Becker, W.; Kölle, S.; Eisenreich, N.; Eyerer, P. *Talanta* **1999**, *50*, 283–290.

[20] DeThomas, F.; Hall, J.; Monfre, S. *Talanta* **1994**, *41*, 425–431.

[21] Wu, Y. Optimization methods for quantitative measurement of glucose based on near-infrared spectroscopy. Ph.D. thesis, The University of Iowa, 2009.

[22] Ingle Jr, J.; Crouch, S. *Spectrochemical Analysis*; Prentice Hall: Englewood Cliffs, NJ, 1988.

[23] Michelson, A. *Philosophical Magazine* **1891**, *31*, 338–346.

[24] Michelson, A. *Philosophical Magazine* **1892**, *34*, 280–299.

[25] Mertz, L. *Transformation in Optics*; Wiley: New York, 1965.

[26] Chase, D. *Applied Spectroscopy* **1982**, *36*, 240–244.

[27] Griffiths, P. *Analytical Chemistry* **1992**, *64*, 868A–875A.

[28] Stuart, B. *Infrared Spectroscopy*; Wiley Online Library, 2004.

[29] Smith, B. *Fundamentals of Fourier Transform Infrared Spectroscopy*; CRC Press: Boca Raton, Florida, 2009.

[30] Kuptsov, A.; Zhizhin, G. *Handbook of Fourier Transform Raman and Infrared Spectra of Polymers*; Elsevier Science: New York, 1998.

[31] Tran, C.; Furlan, R. *Analytical Chemistry* **1992**, *64*, 2775–2782.

[32] Chang, I. *Optical Engineering* **1981**, *20*, 824–829.

[33] Burns, D.; Ciurczak, E. *Handbook of Near-Infrared Analysis*; CRC Press: Boca Raton, FL, 2008.

[34] Tran, C. *Talanta* **1997**, *45*, 237–248.

[35] Xu, J.; Stroud, R. *Acousto-Optic Devices: Principles, Design, and Applications*; Wiley: New York, 1992.

[36] Flanigan, D. *Proceedings of SPIE* **1996**, *2763*, 2–17.

[37] Flanigan, D. *Applied Optics* **1986**, *25*, 4253–4260.

[38] Shaffer, R.; Combs, R. *Software for Generating Synthetic Passive Fourier Transform Infrared Interferograms and Single-Beam Spectra.*; 1999.

[39] Flanigan, D. *Applied Optics* **1997**, *36*, 7027–7036.

[40] Rabiner, L.; Gold, B. *Theory and Application of Digital Signal Processing*; Prentice Hall: Englewood Cliffs, NJ, 1975.

[41] Parks, T.; Burrus, C. *Digital Filter Design*; Wiley: New York, 1987.

[42] Chen, C. *Digital Signal Processing: Spectral Computation and Filter Design*; Oxford University Press, Inc.: New York, 2000.

[43] Small, G.; Harms, A.; Kroutil, R.; Ditillo, J.; Loerop, W. *Analytical Chemistry* **1990**, *62*, 1768–1777.

[44] Mattu, M.; Small, G. *Analytical Chemistry* **1995**, *67*, 2269–2278.

[45] Cingo, N.; Small, G.; Arnold, M. *Vibrational Spectroscopy* **2000**, *23*, 103–117.

[46] Sulub, Y.; Small, G. *Analyst* **2007**, *132*, 330–337.

[47] Bialkowski, S. *Analytical Chemistry* **1989**, *61*, 1308–1310.

[48] Krauss, T.; Shure, L.; Little, J.; MathWorks, I. *Signal Processing Toolbox for Use with MATLAB®: User's Guide*; The MathWorks: Natick, MA, 1994.

[49] Williams, A.; Taylor, F. *Electronic Filter Design Handbook*; McGraw-Hill: New York, 1981; Vol. 198.

[50] Oppenheim, A.; Shafer, R.; Buck, J. *Discrete-Time Signal Processing*; Prentice Hall Press: Englewood Cliffs, NJ, 1989.

[51] Daubechies, I. *IEEE Transactions on Information Theory* **1990**, *36*, 961–1005.

[52] Jensen, A.; la Cour-Harbo, A. *Ripples in Mathematics: the Discrete Wavelet Transform*; Springer Verlag: New York, 2001.

[53] Mallat, S. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1989**, *11*, 674–693.

[54] Antonini, M.; Barlaud, M.; Mathieu, P.; Daubechies, I. *IEEE Transactions on Image Processing* **1992**, *1*, 205–220.

[55] Misiti, M.; Misiti, Y.; Oppenheim, G.; Michel, J. *Matlab Wavelet Toolbox User's Guide, Version 3.*; The Math Works, Inc.: Natick, MA, 2004.

[56] Shao, X.; Cai, W.; Pan, Z. *Chemometrics and Intelligent Laboratory Systems* **1999**, *45*, 249–256.

[57] Chau, F. *Chemometrics: From Basics to Wavelet Transform*; Wiley-Interscience: Hoboken, NJ, 2004.

[58] Alsberg, B.; Woodward, A.; Kell, D. *Chemometrics and Intelligent Laboratory Systems* **1997**, *37*, 215–239.

[59] Wan, B.; Small, G. *Analytica Chimica Acta* **2010**, *681*, 63–70.

[60] Haaland, D.; Easterling, R.; Vopicka, D. *Applied Spectroscopy* **1985**, *39*, 73–84.

[61] Martens, H.; Naes, T. *Multivariate Calibration*; John Wiley & Sons Inc: Chichester, UK, 1992; pp 180–202.

[62] Haaland, D.; Melgaard, D. *Applied Spectroscopy* **2001**, *55*, 1–8.

[63] Haaland, D.; Melgaard, D. *Vibrational Spectroscopy* **2002**, *29*, 171–175.

[64] Haaland, D.; Melgaard, D. *Applied Spectroscopy* **2000**, *54*, 1303–1312.

[65] Melgaard, D.; Haaland, D.; Wehlburg, C. *Applied Spectroscopy* **2002**, *56*, 615–624.

[66] Johnson, R.; Wichern, D. *Applied Multivariate Statistical Analysis*; Prentice Hall: Englewood Cliffs, NJ, 2002.

[67] Brereton, R. *Chemometrics: Data Analysis for the Laboratory and Chemical Plant*; John Wiley & Sons Inc: Chichester, UK, 2003.

[68] *E1655-00: Standard Practices for Infrared Multivariate Quantitative Aaalysis*; ASTM International: West Conshohocken, PA, 2000.

[69] Golub, G.; Reinsch, C. *Numerische Mathematik* **1970**, *14*, 403–420.

[70] Wold, H. *Multivariate Analysis* **1973**, *3*, 383–407.

[71] Malinowski, E. R. *Factor Analysis in Chemistry, 2nd Ed.*; Wiley-Interscience: New York, 1991; Chapter 4.

[72] Haaland, D.; Thomas, E. *Analytical Chemistry* **1988**, *60*, 1193–1202.

[73] Geladi, P.; Kowalski, B. *Analytica Chimica Acta* **1986**, *185*, 1–17.

[74] Kennedy, J.; Eberhart, R. *Proceedings of the IEEE International Conference on Neural Networks* **1995**, *4*, 1942–1948.

[75] Eberhart, R.; Shi, Y. *Proceedings of the 2001 Congress on Evolutionary Computation* **2001**, *1*, 81–86.

[76] Poli, R.; Kennedy, J.; Blackwell, T. *Swarm Intelligence* **2007**, *1*, 33–57.

[77] Goldberg, D. *Genetic Algorithms in Search, Optimization, and Machine Learning*; Addison-Wesley: Reading, MA, 1989.

[78] Shaffer, R.; Small, G. *Analytical Chemistry* **1997**, *69*, 236–242.

[79] Robinson, J.; Rahmat-Samii, Y. *IEEE Transactions on Antennas and Propagation* **2004**, *52*, 397–407.

[80] Wang, K.; Huang, L.; Zhou, C.; Pang, W. *Machine Learning and Cybernetics, 2003 International Conference on* **2003**, *3*, 1583–1585.

[81] Herget, W.; Brasher, J. *Optical Engineering* **1980**, *19*, 508–514.

[82] Xiao, H.; Levine, S.; Nowak, J.; Puskar, M.; Spear, R. *American Industrial Hygiene Association Journal* **1993**, *54*, 545–556.

[83] Shaffer, R.; Combs, R. *Applied Spectroscopy* **2001**, *55*, 1404–1413.

[84] Sulub, Y.; Small, G. *Applied Spectroscopy* **2007**, *61*, 406–413.

[85] Cingo, N.; Small, G. *Applied Spectroscopy* **1999**, *53*, 1556–1566.

[86] Ballard, J.; Remedios, J.; Roscoe, H. *Journal of Quantitative Spectroscopy and Radiative Transfer* **1992**, *48*, 733–741.

[87] Chaffin Jr, C.; Marshall, T. *Proceedings of SPIE* **1998**, *3383*, 113–123.

[88] Lide, D. *CRC Handbook of Chemistry and Physics: A Ready-Reference Book of Chemical and Physical Data*; CRC Press: Boca Raton, FL, 2004.

[89] Sharpe, S.; Johnson, T.; Sams, R.; Chu, P.; Rhoderick, G.; Johnson, P. *Applied Spectroscopy* **2004**, *58*, 1452–1461.

[90] Fang, K.; Wang, Y. *Number-Theoretic Methods in Statistics*; Chapman & Hall: London, 1994.

[91] Wray, S.; Cope, M.; Delpy, D.; Wyatt, J.; Reynolds, E. *Biochimica et Biophysica Acta (BBA)-Bioenergetics* **1988**, *933*, 184–192.

[92] Pollard, V.; Prough, D.; DeMelo, A.; Deyo, D.; Uchida, T.; Stoddart, H. *Anesthesia & Analgesia* **1996**, *82*, 269–277.

[93] Huang, H.; Yu, H.; Xu, H.; Ying, Y. *Journal of Food Engineering* **2008**, *87*, 303–313.

[94] Olesberg, J.; Liu, L.; Van Zee, V.; Arnold, M. *Analytical Chemistry* **2006**, *78*, 215–223.

[95] Arnold, M.; Small, G. *Analytical Chemistry* **2005**, *77*, 5429–5439.

[96] Robinson, M.; Eaton, R.; Haaland, D.; Koepp, G.; Thomas, E.; Stallard, B.; Robinson, P. *Clinical Chemistry* **1992**, *38*, 1618–1622.

[97] Burmeister, J.; Arnold, M.; Small, G. *Diabetes Technology & Therapeutics* **2000**, *2*, 5–16.

[98] Kramer, K.; Small, G. *Applied Spectroscopy* **2007**, *61*, 497–506.

[99] Hazen, K.; Arnold, M.; Small, G. *Applied Spectroscopy* **1994**, *48*, 477–483.

[100] Arnold, M.; Small, G. *Analytical Chemistry* **1990**, *62*, 1457–1464.

[101] Trelea, I. *Information Processing Letters* **2003**, *85*, 317–325.

[102] Hassell, D.; Bowman, E. *Applied Spectroscopy* **1998**, *52*, 18.

[103] Lee, J.; Yoo, C.; Choi, S.; Vanrolleghem, P.; Lee, I. *Chemical Engineering Science* **2004**, *59*, 223–234.

[104] Gendrin, C.; Roggo, Y.; Spiegel, C.; Collet, C. *European Journal of Pharmaceutics and Biopharmaceutics* **2008**, *68*, 828–837.

[105] Chen, J.; Liu, K. *Chemical Engineering Science* **2002**, *57*, 63–75.

[106] Hinz, D. *Analytical and Bioanalytical Chemistry* **2006**, *384*, 1036–1042.

[107] Callis, J.; Illman, D.; Kowalski, B. *Analytical Chemistry* **1987**, *59*, 624–637.

[108] Cregg, J.; Vedvick, T.; Raschke, W. *Nature Biotechnology* **1993**, *11*, 905–910.

[109] Cereghino, J.; Cregg, J. *FEMS Microbiology Reviews* **2000**, *24*, 45–66.

[110] Daly, R.; Hearn, M. *Journal of Molecular Recognition* **2005**, *18*, 119–138.

[111] Kourti, T. *Analytical and Bioanalytical Chemistry* **2006**, *384*, 1043–1048.

[112] Joe Qin, S. *Journal of Chemometrics* **2003**, *17*, 480–502.

[113] Bakshi, B. *AIChE Journal* **1998**, *44*, 1596–1610.

[114] Kateman, G.; Buydens, L. *Quality Control in Analytical Chemistry*; Wiley: New York, 1993.

[115] Haaland, D.; Easterling, R. *Applied Spectroscopy* **1982**, *36*, 665–673.

[116] Sulub, Y.; Small, G. *Analytical Chemistry* **2009**, *81*, 1208–1216.

[117] Small, G.; Arnold, M.; Marquardt, L. *Analytical Chemistry* **1993**, *65*, 3279–3289.

[118] Shaffer, R.; G.W., S.; Arnold, M. *Analytical Chemistry* **1996**, *68*, 2663–2675.

[119] Tsai, F.; Philpot, W. *Remote Sensing of Environment* **1998**, *66*, 41–51.

[120] Becker, B.; Lusch, D.; Qi, J. *Remote Sensing of Environment* **2005**, *97*, 238–248.

[121] Barache, D.; Antoine, J.; Dereppe, J. *Journal of Magnetic Resonance* **1997**, *128*, 1–11.

[122] Serrai, H.; Senhadji, L.; De Certaines, J.; Coatrieux, J. *Journal of Magnetic Resonance* **1997**, *124*, 20–34.

[123] Chau, F.; Shih, T.; Gao, J.; Chan, C. *Applied Spectroscopy* **1996**, *50*, 339–348.

[124] Ehrentreich, F.; Sümmchen, L. *Analytical Chemistry* **2001**, *73*, 4364–4373.

[125] Ma, C.; Shao, X. *Journal of Chemical Information and Computer Sciences* **2004**, *44*, 907–911.

[126] Walczak, B.; van Den Bogaert, B.; Massart, D. *Analytical Chemistry* **1996**, *68*, 1742–1747.